Using large language models for financial advice[*]

Christian Fieberg[†] Lars Hornuf[‡] Maximilian Meiler[§] David J. Streich[¶]

January 30, 2025

Abstract

We study whether large language models (LLMs) can generate suitable financial advice and which LLM features are associated with higher-quality advice. To this end, we elicit portfolio recommendations from 32 LLMs for 64 investor profiles, which differ in their risk preferences, home country, sustainability preferences, gender, and investment experience. Our results suggest that LLMs are generally capable of generating suitable financial advice that takes into account important investor characteristics when determining market and risk exposures. The historical performance of the recommended portfolios is on par with that of professionally managed benchmark portfolios. We also find that foundation models and larger models generate portfolios that are easier to implement and more sensitive to investor characteristics than fine-tuned models and smaller models. Some of our results are consistent with LLMs inheriting human biases such as home bias. We find no evidence of gender-based discrimination, which can be found in human financial advice.

Keywords: Generative AI · artificial intelligence · large language models · financial advice · portfolio management
JEL-Classification: G00 · G11 · G40
Declaration of interest: None

[*] Lars Hornuf and David Streich gratefully acknowledge funding from the German Research Foundation (DFG, project number 552870289). We thank Niklas Töllich and Arnes Triemer for excellent research assistance. We are also grateful to Tom Dudda, Alejandro Lopez-Lira, Matthias Mattusch, Vatsala Shreeti, and Michael Weber, as well as participants of the 8th Luxembourg Workshop on Household Finance and Consumption 2023, the 2024 Journal of Banking & Finance "Generative AI in Finance" paper development workshop in Dresden, and the 2024 "Generative AI in Finance" symposium in Duisburg for their valuable suggestions. All remaining errors are ours.

[†] City University of Applied Sciences, Bremen, Germany. E-Mail: christian.fieberg@hs-bremen.de

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

[§] Dresden University of Technology, Germany. E-Mail: maximilian.meiler@tu-dresden.de

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

1 Introduction

Financial advice has the potential to improve financial outcomes by increasing stock market participation and by mitigating biases such as under-diversification (Calvet et al., 2009; Gennaioli et al., 2015). However, access to financial advice has thus far been limited to wealthy investors due to substantial advisory fees and minimum investment amounts (Philippon, 2016). This implies that less wealthy investors, who stand to benefit most from such advice (Bhattacharya et al., 2012), have faced the greatest barriers to accessing it. Recently, digital financial advice platforms, often referred to as "robo-advisors", have entered the market as low-cost, technology-based alternatives to individual financial advisors (Jung et al., 2018; Rühr et al., 2019).

Robo-advisors use standardized online risk questionnaires and rule-based allocation algorithms to provide financial advice and delegated investment, which eliminates the impact of individual advisors' idiosyncrasies (Rossi and Utkus, 2024; Foerster et al., 2017). Nevertheless, while robo-advisors can reduce biases (D'Acunto et al., 2019), decision-makers have demonstrated reluctance to follow algorithmic advice (Dietvorst et al., 2015).[1] A rule-based advice system also cannot fully eliminate the conflict of interest inherent in the advisor–investor relationship, which can result in self-serving advice (Christoffersen et al., 2013; Chalmers and Reuter, 2020) and underperformance of passive benchmarks (Bergstresser et al., 2009; Stolper and Walter, 2017).

The present paper investigates whether large language models (LLMs) such as ChatGPT are a suitable source of financial advice. With rapid improvements in the capabilities of LLMs (the most prevalent form of generative artificial intelligence (AI)), their disruptive potential is becoming readily apparent. Firms in industries as diverse as education, translation services, and software are forced to adjust their business models in light of increasingly capable AI tools (The Economist, 2024b). The financial advice domain presents a compelling application of LLMs for two reasons. First, LLMs may accelerate the democratization of access to financial advice through digital technology. Specifically, LLMs may provide substantial improvements in the quality of digital financial advice due to their conversational capability (cf. Lo and Ross, 2024), which may help further reduce the required human interaction. LLMs are able to elaborate on the rationale behind a specific recommendation, which may increase adoption (Litterscheidt and Streich, 2020). This elaboration may be particularly relevant for algorithm-averse decision-makers (Dietvorst et al., 2015), whose concerns may be alleviated through the algorithm's ability to learn from previous mistakes (Berger et al., 2021). LLMs are also capable of simulating a personal relationship with the investor (Safdari et al., 2023), which the financial advice literature considers an important aspect of advice provision (Gennaioli et al., 2015; Germann et al., 2025). Stolper and Walter (2019), for example, show that when

[1] While Germann and Merkle (2023) find no evidence of general algorithm aversion in an experimental financial decision-making context, 44% of participants in their study initially chose to delegate to a human rather than an algorithmic decision support system, which suggests that some investors are algorithm-averse.

advisor and client share common demographic features, clients are more likely to follow the advisor's recommendation.

Second, early tests of the capabilities of LLMs in the finance domain suggest that they may be well-suited to provide financial advice. LLMs are able to reproduce the correct answers to standard financial literacy questions (Niszczota and Abbas, 2023) and financial licensing exam preparation questions (Fairhurst and Greene, 2025). They can also extract sentiment from text to forecast stock prices (Lopez-Lira and Tang, 2023; Kim et al., 2023a), firm-level investments (Jha et al., 2024), and even macroeconomic shocks (Hansen and Kazinnik, 2023). Unlike human investors, who can be overwhelmed by large amounts of financial information (Hirshleifer et al., 2009; Frederickson and Zolotoy, 2016; Schmidt, 2019), LLMs may be able to efficiently distill relevant information from financial texts. Furthermore, their capabilities are not limited to textual input data. Korinek (2023), who details use cases of LLMs for economic research, uploaded historical stock market prices for three stocks to ChatGPT's Advanced Data Analysis tool and had it successfully compute the stocks' market betas and other portfolio metrics. Thus, state-of-the-art LLMs seem capable of employing quantitative analysis required for portfolio management applications. The potential of LLMs in the financial industry is evident from their early adoption by leading asset managers and financial service providers. For example, the asset management division of Morgan Stanley, a major American bank, partnered early with OpenAI to develop a model trained on the bank's internal data (Reuters, 2023). Bloomberg, a provider of financial data and software, has recently released an LLM trained for natural language processing tasks using a vast dataset of high-quality financial text (Wu et al., 2023).

Against the backdrop of these developments, this study raises two research questions. First, can current LLMs generate suitable financial advice? And second, which LLM features are associated with higher-quality financial advice? To address these questions, we construct 64 hypothetical investor profiles differing with respect to their risk tolerance and risk capacity, sustainability preferences, gender, investment experience, and home country. The profiles capture key dimensions financial advisors are required to take into account when generating portfolio recommendations (ESMA, 2018; SEC, 2019). To assess the ability of LLMs to provide financial advice to international investors, we consider Chinese, German and US investor profiles to reflect the largest economies in Asia, Europe and the Americas. For each of the 64 profiles, we elicit portfolio recommendations from 32 state-of-the-art LLMs, which we categorize according to their type (foundation models vs. fine-tuned models), size (as measured by the number of parameters), and license (proprietary vs. open-source). *Foundation models* are pre-trained, general-purpose models such as OpenAI's GPT-4 or Google's Gemini. *Fine-tuned models* employ machine learning techniques to alter existing foundation models' weights for a specific purpose (e.g., solving specific tasks or applying improved reasoning capabilities).[2] We use the number of parameters used in the training of an LLM as a measure of its complexity and size. Larger LLMs have been shown to learn and retain more complex relationships within the training data, which allows them to produce higher-quality output with higher generalizability than smaller LLMs (Ding et al., 2023; Brown et al., 2020). For LLMs published under an *open-source license*, information on the training process, input data, and LLM configuration are made available to developers and researchers. Developers of *proprietary models*, on the other hand, do not publish details on their LLMs.

Based on the results of early capability tests, we hypothesize that LLMs should generally be able to provide financial advice due to the information contained in commonly used training data. We expect foundation models to be better suited to provide financial advice due to the risks associated with using fine-tuned models outside their specific domains. We further expect more complex LLMs (as measured by the number of parameters) to be better suited to provide financial advice given the higher generalizability of output and scaling properties observed in other tasks.

To test our hypotheses, we investigate three distinct suitability dimensions for 2,048 portfolio recommendations (64 profiles × 32 LLMs). First, we are interested in the extent to which the recommended portfolios can in fact be implemented by retail investors. We use the number of suggested portfolio assets, the share of assets with publicly available price data, and the incidence of an array of response errors (e.g., non-existent ticker) to measure the implementability of portfolios. We consider portfolios with an excessively high number of assets, low degree of data availability, and high incidence of response errors to be more difficult to implement.

Second, we are interested in the extent to which the portfolios' exposure to asset classes, markets, and various risk components is in line with the investor profiles' characteristics. While there is a lack of consensus on what constitutes an optimal portfolio, we use the portfolio exposures reported in related financial advice studies as a benchmark (Foerster et al., 2017; Bhattacharya et al., 2012, 2024; Rossi and Utkus, 2024; Scherer and Lehner, 2023). In addition, we compare the exposures of the LLM-generated recommendations to those recommendations we elicit from 20 large robo-advisors operating in Germany and the United States.

Third, we are interested in the historical risk-adjusted performance of the recommended portfolios. In addition to simple risk-adjusted performance measures (excess returns and Sharpe ratios), we compute alphas to the Fama–French six-factor model (FF6) (cf. Fama and French, 1993; Carhart, 1997; Fama and French, 2018). To account for operating expenses, we additionally include the portfolios' total expense ratio (TER) in our analyses. We compare the

[2] While there are LLMs that are fine-tuned for the financial context (Wu et al., 2023; Yang et al., 2023), none of the LLMs we consider are fine-tuned to the financial context in general, nor to the financial advice context specifically.

performance of LLM-generated recommendations to the professionally managed benchmark portfolios obtained from the robo-advisors.[3]

We present five main results. First, LLMs are generally able to recommend portfolios that can in fact be implemented. Portfolio recommendations include specific securities (mostly low-cost ETFs), as well as specific portfolio weights and exchange tickers. While minor response errors do occur, they are concentrated in a few LLMs, and error incidence can be substantially reduced through prompt engineering.

Second, the exposure in LLM-generated portfolio recommendations is in line with the prescriptions of modern portfolio theory, as well as the reported exposures in portfolios managed by human financial advisors (Foerster et al., 2017; Bhattacharya et al., 2024; Jacobs et al., 2014) and robo-advisors (Rossi and Utkus, 2024; Scherer and Lehner, 2023). Specifically, the correlation of investor characteristics and exposure variables is in line with sensible advice. Among all investor characteristics, risk tolerance is by far the greatest determinant of portfolio exposure. Moreover, exposure in LLM-generated portfolio recommendations is determined to a far lesser extent by advisor fixed effects than in portfolios recommended by the robo-advisors in our sample or in the portfolios recommended by human advisors in Foerster et al.'s (2017) sample. This result suggests that LLM-generated advice is more consistent than advice given by human advisors and robo-advisors.

Third, the historical risk-adjusted performance of LLM-generated portfolio recommendations is no worse than that of professionally managed robo-advisory portfolios. Specifically, Sharpe ratios and excess returns are significantly higher for LLM-recommended portfolios than robo-advisory portfolios, while six-factor alphas are significantly lower. This pattern suggests that performance in LLM portfolios is achieved to a greater extent by exposure to commonly used asset pricing factors. Addressing concerns over look-ahead bias in our performance estimates, we find little evidence that recommendations by LLMs with more recent access to information generate higher performance. The concerns are further alleviated by recent descriptive evidence regarding LLM performance following a GPT-4 information update.

Fourth, we find evidence that foundation models and larger LLMs provide portfolio recommendations that are easier to implement than those of fine-tuned models and smaller LLMs due to a lower likelihood of response errors. Our results also suggest that foundation models and larger models recommend portfolios that are better suited to individual investor characteristics. Specifically, foundation models and larger LLMs are more sensitive to an investor's risk tolerance when determining exposure to equity and risk. We find no systematic differences

[3] Given that our performance measures may suffer from look-ahead bias (i.e., LLMs recommending portfolios that have historically done well), they may overestimate the potential performance of LLM-generated portfolio recommendations. We argue that, because we do not specify a particular historical period and because we compare the performance of LLM-generated recommendations to that of professionally managed portfolios (that also have access to historical return patterns), our results are less affected by look-ahead bias than the results of specific, short-window forecasting tasks. Nonetheless, the performance results should be interpreted with caution.

in the risk-adjusted returns of foundation versus fine-tuned models or larger versus smaller LLMs.

Fifth, we find evidence that LLMs inherit some well-established biases from their training data. Specifically, we find that on average, LLM recommendations suffer from home bias, i.e., excessive allocation to domestic securities (Coval and Moskowitz, 1999). In contrast, the robo-advisory benchmark portfolios do not suffer from home bias. We find no evidence of gender-based discrimination in the exposure of the portfolios as has been documented in a recent audit study involving human advisors (Bhattacharya et al., 2024).

We contribute to the literature on the suitability of human (Foerster et al., 2017; Linnainmaa et al., 2021; Bhattacharya et al., 2012, 2024) and digital financial advice (D'Acunto et al., 2019; Rossi and Utkus, 2024; Scherer and Lehner, 2025) by expanding the limited body of evidence suggesting that off-the-shelf LLMs may in fact be used for financial advice. Pelster and Val (2024) demonstrate that ChatGPT may be useful for picking stocks based on a real-time experiment around corporate earnings announcements. Fieberg et al. (2023), Oehler and Horn (2024), and Hens and Nordlie (2024) provide some preliminary evidence that ChatGPT can generate suitable portfolio recommendations. We are the first study to elicit recommendations from a large number of diverse LLMs (32) and for a large number of different investor profiles (64), which allows for a representative assessment of the capabilities of LLMs in the financial advice domain. It also allows us to uncover systematic differences in the quality of financial advice according to LLM features such as size and type. While we find that LLM-generated recommendations are sound, further research should explore what drives user acceptance of these recommendations.

We also contribute to research investigating the capabilities of LLMs in various domains. Such studies typically investigate questions with objectively correct answers such as multiple-choice financial literacy questions (Niszczota and Abbas, 2023) or accounting certification exams (Eulerich et al., 2024). In our study, we investigate the performance of LLMs in a complex applied problem with no single correct answer — there is no consensus on what constitutes "correct" financial advice (Lo and Foerster, 2023). Thus, we do not examine whether LLMs are able to reproduce finance-related knowledge, but rather whether they can apply it appropriately, which is likely what will determine the disruptive potential of LLMs in the financial sector.

The remainder of the paper is organized as follows. Section 2 derives hypotheses based on related literature on LLM capabilities and requirements for financial advice. Section 3 details how we elicit portfolio recommendations and construct the suitability measures. Section 4 presents the empirical results. Section 5 concludes and provides an outlook on the use of LLMs in the financial advice context.

2 Theoretical framework

2.1 Can LLMs generate suitable financial advice?

At least two conditions must be met for LLMs to generate suitable financial advice. LLMs must have (i) access to a specific set of information in their training data and (ii) the capability to apply this information in generating financial advice.

Specifically, LLMs must have access to information in three domains. First, they must have access to basic financial theory (e.g., Markowitz, 1952; Tobin, 1958). Knowledge of these basic principles ensures that LLMs internalize the normative optimum of a passive investment strategy, as well as the relationship between investor characteristics such as risk preferences and the optimal portfolio. Second, LLMs must have access to a range of specific financial products as well as some details on their exposure and risk-return profiles. If LLMs cannot draw on a specific set of real funds, stocks, or bonds, they will either be unable to recommend a portfolio of specific securities or else "hallucinate," i.e., recommend non-existent or incorrectly specified products (Huang et al., 2023). Third, LLMs must have access to recent market and company news in order to adjust portfolio suggestions to market and idiosyncratic risk environments (cf. Cready and Gurun, 2010).

To assess *ex ante* whether LLMs are likely to have access to the information detailed above, we analyze the domains contained in the CommonCrawl dataset, which most LLMs use as at least part of their training dataset (Zhao et al., 2023).[4] CommonCrawl is one of the largest publicly available datasets and has been compiled through regular scrapes of internet content since 2008. Table A1 in the Appendix lists domains contained in CommonCrawl that provide suitable information on basic financial theory (educational domains, as well as general-purpose and finance-specific knowledge domains), specific investment products (financial analysis platforms such as TradingView), and time-stamped news articles (e.g., Euronews). Thus, CommonCrawl, which is part of most LLMs' training data, contains information on all three information domains deemed necessary to provide suitable financial advice.

In addition to having access to this information, LLMs must also have the capability to correctly interpret the information to form recommendations. Regarding basic financial concepts, Niszczota and Abbas (2023) assess the ability of GPT-3.5 and GPT-4 to answer a set of 21 single-choice financial literacy questions (Mitchell and Lusardi, 2022; Heinberg et al., 2014). The results suggest that advanced LLMs are financially literate; while GPT-3.5 was able to answer approximately two-thirds of the questions correctly, GPT-4 obtained a near-perfect score. As a benchmark, data from the 2019 Survey of Consumer Finances suggest

[4] A list of domains contained in the full dataset is available online: https://commoncrawl.github.io/cc-crawl-statistics/plots/mimetypes.

that fewer than half of US respondents were able to answer the "Big Three" financial literacy questions correctly (Lusardi and Mitchell, 2023).[5]

Regarding information on specific investment products, Fieberg et al. (2023) document that GPT-4 is indeed able to produce suitable portfolio suggestions containing specific tickers and security names when prompted for financial advice. Korinek (2023) finds that ChatGPT's Advanced Data Analysis tool is able to compute market betas from historical price data and illustrate evolving portfolio weights over time.

There is also ample evidence that LLMs are well-suited to process financial news and extract the relevant implications for portfolio management purposes. Lopez-Lira and Tang (2023) use ChatGPT to classify the sentiment in news article headlines and show that subsequent risk-adjusted stock returns correlate significantly with the classifications. LLMs are also able to synthesize the most important pieces of information from corporate disclosures (Kim et al., 2023a). The findings of these studies are corroborated by the strong performance of LLMs (mostly of OpenAI's GPT series) in assessing idiosyncratic risk (Kim et al., 2023b) and unusual communication (Beckmann et al., 2024) from earnings call transcripts, as well as monetary policy expectations from central bank communications (Cook et al., 2023). Addressing concerns over look-ahead bias in LLM predictions, Pelster and Val (2024) have run a live experiment to show that ChatGPT can provide valuable out-of-sample stock price forecasts from firm-level information.

Given that the relevant information is likely contained in most LLMs' training data, and given that early studies document strong capabilities in reproducing basic financial concepts, using financial data for portfolio management purposes, and interpreting financial news, we hypothesize:

Hypothesis 1. *LLMs are capable of generating suitable financial advice.*

2.2 Which LLM features are associated with better financial advice?

Within our theoretical framework, specific LLM features may affect the quality of advice by determining how a set of training data is used to generate financial advice (see Figure 1). Specifically, we investigate differences arising from the type of model (foundation model vs. fine-tuned model) and its complexity (number of parameters).[6]

[Figure 1 approximately here]

[5] While these results suggest that LLMs are exposed to key financial concepts in their training data and are increasingly able to reproduce these concepts, this does not ensure that LLMs are able to apply financial knowledge. For example, Smith (2024) shows that GPT-4 ignores the time value of money when prompted for advice on a car loan.

[6] When investigating LLM performance, we argue that the information cut-off may affect the quality of an LLM's training data. We expect models with regular information updates or more recent information cut-offs to have more recent information on financial news and prior returns, which should improve performance.

We distinguish between *foundation models*, which are pre-trained, general-purpose models such as OpenAI's GPT-4 or Google's Gemini, and *fine-tuned models*, which adjust existing foundation models using a narrower "downstream" dataset to suit a specific purpose (e.g., to converse, reason, or make financial predictions) through machine learning techniques (Ovadia et al., 2023). The fine-tuned models in our sample are mostly re-trained to improve conversational skills. While fine-tuning can improve an LLM's performance in a specific task, there are risks associated with it, particularly when these LLMs are used outside the domains they have been re-trained for, as is the case with the LLMs we use for this study.[7] Excessive fine-tuning can lead to poor performance outside the intended domain due to over-fitting (overly adjusting weights to the downstream dataset) and "catastrophic forgetting" (omission of previously learned relationships due to parameter updates during the fine-tuning stage) (Luo et al., 2023).[8] Fine-tuning may also increase the likelihood of hallucination when there is a discrepancy between the original training data and the downstream dataset (capability misalignment) or when LLMs are trained to provide answers that appease human evaluators (belief misalignment) (cf. Zhao et al., 2023; Huang et al., 2023). Given these risks, we expect foundation models to be better suited to provide financial advice:

Hypothesis 2. *Foundation models are better suited to provide financial advice than fine-tuned models.*

We measure LLM *complexity* using the number of parameters an LLM employs. LLMs with more parameters are able to learn and retain more complex relationships within the training data, which allows them to perform tasks they were not explicitly trained on. Larger LLMs tend to produce higher-quality output with higher generalizability than smaller LLMs (Brown et al., 2020). Ding et al. (2023) find that the performance gap between large and small LLMs is greater for few-shot evaluations than for zero-shot or one-shot evaluations, suggesting that larger LLMs are better able to meta-learn from tasks. In the finance domain, Lopez-Lira and Tang (2023) demonstrate that greater LLM complexity enhances the capability to anticipate future stock price movements from news. They show that larger LLMs (GPT-4) outperform smaller LLMs (GPT-1, GPT-2), especially for complex news articles. While the LLMs do not only differ with respect to their size, the results suggest that predicting returns from news headlines is an emerging capability of larger LLMs. We thus expect larger models to be better suited to provide financial advice.

Hypothesis 3. *Larger LLMs are better suited to provide financial advice than smaller LLMs.*

[7] While there are some LLMs fine-tuned for interpreting financial text (Wu et al., 2023), we were unable to collect data for these models. Besides, there is some evidence that large LLMs perform as well in financial text interpretation as LLMs narrowly trained on this task (Kim et al., 2024; Li et al., 2023).

[8] Approaches to mitigate this include randomly replacing the fine-tuned weights with their pre-trained predecessors, thus limiting the impact of the fine-tuning on model weights (Luo et al., 2023).

Finally, we distinguish between *open-source* LLMs, for which code and training data are made available to the developer and research communities, and *proprietary* LLMs, whose developers do not publish code or model details such as the LLM's size. *A priori*, it is unclear whether open-source or proprietary LLMs perform better. On the one hand, developers of LLMs with a higher performance and thus profit outlook may self-select into proprietary licenses to protect the competitive advantage of their LLMs or maintain effort incentives (cf. August et al., 2021). On the other hand, open-source LLMs benefit from third-party contributors and increased public scrutiny, which may help identify weaknesses and foster trust (The Economist, 2024c). Thus, we do not formulate an explicit hypothesis regarding an LLM's license type.

3 Method

This section details our choices in defining investor profiles, formulating standardized prompts to elicit recommendations, compiling a list of recent LLMs suitable for our assessment, obtaining recommendations from robo-advisors to serve as a benchmark, and constructing the various suitability measures for our analyses.

3.1 Investor profiles

Regulatory guidelines on the provision of financial advice (e.g., MiFID II) stipulate that financial advisors must take into account their clients' individual circumstances when providing financial recommendations. We thus construct hypothetical investor profiles to assess whether LLMs tailor their recommendations to the individual investor. We define 64 investor profiles which vary with respect to their risk capacity, risk tolerance, home country, sustainability preferences, gender, and investment experience (see Table A2 in the Appendix).

According to modern portfolio theory, the optimal portfolio is determined solely by an individual's risk preferences. Previous research has distinguished between an investor's level of *risk capacity*, which is related to future financial obligations that impose a constraint on the investor's liquidity, and *risk tolerance*, which reflects an individual's inherent willingness to take risks and is not limited to the financial domain (Frey et al., 2017; ESMA, 2018; Streich, 2023). Bhattacharya et al. (2012) elicit risk preferences by asking investors to select one of six risk categories and measure risk capacity through demographic data and wealth proxies. In line with this, we capture differences in risk capacity by varying our investor profiles' age and investment horizon. Typically, as the investment horizon shortens, investors will shift portfolios toward less volatile securities to account for the liquidity constraint imposed by the impending conclusion of the investment period. We capture differences in risk tolerance by varying our investor profiles' subjective risk tolerance.

To assess whether the LLMs take into account the investor's home country, we vary the investor profiles' home countries. There are a number of ways an investor's home country may affect the suggested portfolios. On the one hand, investments in domestic securities eliminate exchange rate risk and and could provide investors with better access to information (Aghion and Bolton, 1992; Hornuf et al., 2022). On the other hand, excessively investing in domestic securities creates cluster risk as both capital and labor income are affected by the economic situation in the investor's home country (French and Poterba, 1991; Coval and Moskowitz, 1999; Ivković and Weisbenner, 2005).

To assess whether LLM-generated investment advice takes into account investors' sustainability preferences, we include some profiles with an explicit sustainability preference. Regulatory frameworks stipulate that financial advisors should "[...] identify an investment strategy that fulfills the client's sustainability preferences [...]" (ESMA, 2023, p. 23). In practice, this is usually achieved by investing in funds that explicitly exclude securities of firms whose business model is inconsistent with sustainability in either of the environmental, social, or governance (ESG) pillars.

We also include some investor profiles that explicitly reference the investor's gender. While gender itself should not affect portfolio recommendations, the household finance literature consistently documents gender differences in risk-taking propensities and stock market participation (Croson and Gneezy, 2009; Halko et al., 2012). Besides, human advisors have been shown to discriminate based on gender (Bhattacharya et al., 2024). If the LLMs, which are trained on text produced by humans, inherit these gender biases, they might adjust their portfolio suggestions according to the investor's gender.

Finally, we explicitly vary the level of an investor's experience in some profiles. Financial advisors are typically required to ensure their clients understand the nature of the risks associated with the investment products they are recommended. In addition, there is some evidence that seemingly more experienced investors receive higher-quality financial advice (Bucher-Koenen and Koenen, 2015).

Table A2 in the Appendix lists all investor profiles used in the analysis, each representing different combinations of age and investment horizon, risk tolerance, home country, sustainability preferences, gender, and experience.

3.2 Requests

To obtain portfolio recommendations, we formulate standardized requests for suitable portfolio allocations for each of the investor profiles. Since most LLMs are not designed to provide individual investment advice upon request (likely for legal reasons), requests need to be formulated in a hypothetical scenario to obtain specific portfolio suggestions (Lo and Ross, 2024). We use the following prompt:

> I am [gender], [age] years old and live in [home country]. I am investing for retirement over a [investment horizon]-year horizon, and have a [risk tolerance] risk tolerance. I am an [experience] investor. Environmental aspects are [not] important to me when I am investing. Which specific financial products (including ticker

> and provider) would a typical financial advisor recommend for investment given my circumstances? Which composition (as a percentage) would he recommend for each financial product? I will not consider your response personalized advice.

Each placeholder is replaced with the respective profile's characteristics as presented in Table A2 in the Appendix. Data was collected between January 18, 2024, and January 23, 2024.[9] Portfolio recommendations, including security ticker and portfolio weights, are compiled for each LLM–profile combination. As an illustration, Figure 2 displays GPT-3.5's response to an exemplary request. There is a lot to like about the response. First, it suitably recommends a high share of equities for this specific profile given the high risk tolerance and long investment horizon. Second, the specific products suggested are all low-cost exchange-traded funds (ETFs) managed by well-known asset managers (Vanguard and BlackRock's ETF platform iShares). Third, the response acknowledges the profile's sustainability preferences and recommends an ETF of companies with lower carbon exposure.

Related studies on the performance of LLMs use various prompting techniques such as *masking* (anonymizing company identifiers) to avoid look-ahead bias, i.e., information leakage from the training data biasing in-sample prediction tasks (Pelster and Val, 2024; Sarkar and Vafa, 2024; Alonso, 2024). In our context, look-ahead bias presents a challenge for the analyses of financial performance. Since we are interested in the long-term performance of the recommended portfolios, our performance analyses are by necessity backward-looking.[10] The concern is that LLMs can observe historical returns in their training data, which they could use to recommend optimal portfolios in hindsight. However, the lack of a specific investment period makes *ex post* optimization less likely. To address this concern, we follow the reasoning of Sarkar and Vafa (2024) and others (e.g., Lopez-Lira and Tang, 2023) and use the timing of an LLM's cutoff from information in two additional analyses. First, we account for an LLM's information cutoff date when we investigate the relationship between LLM characteristics and performance. Second, we exploit an update to GPT-4 as a case study to test for potential look-ahead bias in the portfolio recommendations. We discuss the results and implications of these analyses in section 4.3.

[Figure 2 approximately here]

[9] Each prompt is only requested once. To ensure consistent responses, we restrict our analyses to LLMs for which the temperature (i.e., the LLM's tendency toward creativity or randomness) could be set to zero, which ensures that an LLM provides deterministic answers.

[10] A true out-of-sample test would require waiting for 30 years to evaluate the performance of portfolios recommended today.

3.3 Large language models

We study 32 LLMs in our analysis. We initially compiled a list of 76 LLMs featured on common industry leaderboards[11] and added 3 recent and highly publicized releases that were not listed on the leaderboards at the time our list was compiled (Mistral Medium, Gemini, Grok-1). Whenever there are multiple different versions of a particular LLM, we use the largest (in terms of parameters) and most recent version.[12] For example, we use LLaMA2-70B, which draws on 70 billion parameters, instead of the smaller versions LLaMA2-13B or LLaMA2-7B. Of the 79 LLMs, we were unable to access 41.[13] Finally, we restrict our analysis to LLMs whose temperature can be reduced to or is by default zero and whose P parameter can be set to or is by default 1. The former ensures that answers are consistent for identical requests, while the latter ensures that answers are generated from the entire distribution of likely completion tokens. The temperature could not be set to zero for 6 LLMs, leaving us with 32 LLMs for which we obtain portfolio recommendations.

Table 1 contains the LLMs we use for our analyses, as well as some key characteristics. We characterize LLMs according to their type (foundation vs. fine-tuned), size (in billion parameters), and license (proprietary vs. open-source). We define a dummy variable which takes the value one if an LLM uses more than 60 billion parameters.[14] Defining more granular size categories (e.g., tertile or quartile split) leads to collinearity in LLM characteristics. As Table 1 shows, the eight largest LLMs are all foundation models published under a proprietary license, which implies that there is no variation in LLM type among the largest quartile of LLMs.

We further document the date at which the LLMs were cut off from information, which typically coincides with the date the LLM was trained. Given that our performance analyses will be backward-looking, this allows us to assess whether LLMs drawing on more recent information provide better portfolio recommendations, given that they observe a larger share of past return realizations.

Overall, 18 of the 32 LLMs we consider are foundation models. Of the 14 fine-tuned models, 11 LLMs are based at least partially on models from Meta's LLaMA family, likely owing to the open-source nature and high performance of these LLMs. While most LLMs (22) are published under an open-source license (typically Apache 2.0, MIT, or CC BY-NC

[11] We compile a list of all LLMs listed on the LMSYS, Stanford, and Alpaca leaderboards, as well as the top 10 LLMs by performance average from the HuggingFace leaderboard. The LLM list was compiled in December 2023.

[12] The only exception to this is Falcon, where we use Falcon-40B instead of the larger Falcon-180B, for which we could not obtain access.

[13] The majority of these LLMs were outdated versions still featured on the leaderboards. For 5 LLMs, we were unable to establish an API endpoint.

[14] This is effectively a median split (15 large LLMs, 17 small LLMs). As a robustness test, we also employ continuous size specifications.

licenses), some relevant LLMs, including Google DeepMind's Gemini and OpenAI's GPT-3.5 Turbo and GPT-4 Turbo, are not published open-source. Most LLMs employ between 7 and 70 billion parameters, while some LLMs such as the Gemini and presumably the GPT models draw on substantially more parameters. While most LLMs are cut off from information at the date of their initial programming or the latest update, a few LLMs are able to draw on up-to-date online information by regularly re-training the LLMs using recent data. Complexity AI, for example, has developed LLMs based on the Mistral and LLaMA2 models that are able to incorporate recently published online information in their responses.

Table A3 in the Appendix reports pairwise correlations between LLM characteristics. The correlations imply that foundation LLMs and large LLMs are significantly less likely to be published under an open-source license. This may be due to the high development costs of large foundation models (e.g., OpenAI's GPT series, Google's Gemini), which developers would not be willing to incur if their LLMs could be easily replicated by competitors.

[Table 1 approximately here]

3.4 Robo-advisor recommendations

As a benchmark for the LLM-generated portfolio recommendations, we obtain portfolio recommendations for a subset of investor profiles from professional robo-advisors based in Germany and the US. We use robo-advisors as a benchmark for two reasons. First, robo-advisors use an online onboarding process that collects investor characteristics in a structured way and provide explicit portfolio recommendations based on these characteristics. Thus, we can collect specific recommendations from professional financial advisory firms for each of our investor profiles (cf. Scherer and Lehner, 2025). Second, several studies suggest that robo-advisors can improve diversification, reduce biases (D'Acunto et al., 2019), and even outperform human financial advisors (Helms et al., 2022; Rossi and Utkus, 2024; Tao et al., 2021).

We focus our data collection on the 12 standard investor profiles for US and German investors. The profiles vary in risk tolerance, age, and sustainability preferences (profiles 13 through 36; see Table A2 in the Appendix).[15] For both countries, we obtain recommendations from the ten largest robo-advisors by assets under management (AuM) for which recommendations are publicly available.[16] To identify the largest robo-advisors in each country, we compile a list of eligible robo-advisors from industry rankings (Barron's, 2024; Forbes, 2024;

[15] Because all robo-advisors in our sample elicit the investor's experience level, we collect recommendations for the highest and lowest experience level available. In most of our analyses, we compare LLM-generated recommendations to the average of high- and low-experience recommendations. More experienced investors are recommended slightly higher equity shares (62% vs. 56%, $p < 0.05$) and lower fixed-income shares (35% vs. 41%, $p < 0.05$; see Table A5 in the Appendix).

[16] Some robo-advisors only display asset class breakdowns of their recommendations. The exact portfolio allocation including securities is only revealed to actual clients. For other advisors (particularly US advisors), a valid social security or US bank account number had to be specified in order to obtain a portfolio recommendation.

Morningstar, 2023; ExtraETF, 2023) and related academic studies (Helms et al., 2022; Oehler and Horn, 2024; Scherer and Lehner, 2023) and collect AuM data from their most recent SEC filings (form ADV) or industry reports.[17] We were unable to obtain portfolio recommendations from Chinese robo-advisors due to limited information availability and geographical restrictions. Table A4 provides an overview of the robo-advisors included in our analysis and some of their key characteristics. The list includes digital offerings by leading financial service providers (e.g., Vanguard, Fidelity, and Schwab) as well as dedicated robo-advisory firms (e.g., Scalable and Wealthfront). The robo-advisors in our sample account for approximately two-thirds of the German robo-advisory market and one-fifth of the US robo-advisory market.[18]

We obtain portfolio recommendations using a standardized approach that aims to ensure consistency with the investor profiles used for the LLMs. Since all robo-advisors use more than two risk profiles (see Table A4 in the Appendix), we use the most (least) risk-seeking profile for high (low) risk tolerance profiles. Following Oehler and Horn (2024), we set the initial investment amount at $10,000 (€10,000), with no recurring investments, and disposable monthly net income at $1,500 (€1,500). We provide a detailed account of the screening questions and the corresponding answers in Table A6 in the Appendix. To ensure comparability with German robo-advisors, we do not consider US-specific employee retirement plans, such as 401(k) plans governed by the Employee Retirement Income Security Act (ERISA).

3.5 Construction of suitability measures

3.5.1 Implementability

We construct three measures of implementability: the number of suggested assets per portfolio, the share of assets for which price data is publicly available to retail investors using the provided tickers, and the degree to which the LLMs produce response errors. Portfolios with excessively large numbers of assets are harder for retail investors with a limited portfolio value to implement and re-balance. If no price data is available for an asset, investors cannot use historical data to form expectations regarding the expected return and volatility of specific securities. Because private investors do not have access to commercial databases, we document whether the corresponding price data is available on *YahooFinance*.[19] Finally, if the portfolio recommendations are incomplete or otherwise erroneous, they cannot be easily implemented. We identify five types of potential response errors sorted from most likely to least likely (see Table 2):

[17] The list was compiled in October 2024.

[18] https://www.statista.com/outlook/fmo/fintech/digital-investment/robo-advisors

[19] We use *YahooFinance* for the data availability measure as it is one of the most popular sources of financial data for retail investors. For the six-factor models we use in the exposure and performance analyses, we use time series data from Bloomberg.

1. Portfolio weights do not add up to 100%.
   Example: Openchat-3.5 recommends investing 100% in a "Green Bond ETF," 60% in a "Sustainable Equity ETF," 40% in a "Green Money Market Fund," 10% in a "Green Real Estate ETF" and 10% in a "Green Corporate Bond ETF."

2. No recommendation is generated.
   Example: Falcon-40B-Instruct suggests to "diversify," "consider target-date funds," "invest in low-cost index funds," "consider annuities," and "consider working with a financial advisor," but does not provide a specific recommendation.

3. No portfolio weights are provided.

4. The ticker provided by the LLM is wrong (i.e., does not exist or does not match the suggested product).
   Example: Claude 2.0 recommends allocating 15% to Allianz Germany Green Technology Fund (ticker: ALLI-GRÜN).

5. No ticker is provided.

3.5.2 Time series and six-factor model

We rely on a time series of portfolio values for some of our exposure and performance measures. This subsection describes in detail how we obtain time series from the recommendations generated by the LLMs and robo-advisors. First, we verify the provided ticker. In case there is a minor deviation, we manually correct the ticker (e.g., "BMW" provided, corrected to "BMW.DE"). Second, we make assumptions regarding the portfolio weights of the individual products. If no portfolio weights are provided by the LLM, even though specifically requested, we assume equal weighting of all suggested products. If ranges are provided for the weights, we use the average of the lower and upper bound as a portfolio weight. Finally, we divide all portfolio weights by the total portfolio weight to account for cases in which the (adjusted) weights in a portfolio do not add up to 100%.

Next, we obtain monthly price data from January 2010 to December 2023 from Bloomberg. We use the adjusted closing price series, which is quoted in US dollars and accounts for stock splits and dividends. Portfolio values are constructed using the suggested portfolio weights and assuming monthly rebalancing.[20]

[20] Missing product or portfolio returns are replaced with zero-returns. We assume annual rebalancing in robustness tests.

To obtain our measures of risk-adjusted returns, exposure to market risk, and idiosyncratic volatility, we employ a six-factor regression model of the following form:

$$\begin{aligned} r_{mpt} - r_{ft} = \alpha_{mp} + \beta_{mp}^{mkt}(R_{mkt,t} - r_{ft}) + \beta_{mp}^{SMB} \times SMB_t + \beta_{mp}^{HML} \times HML_t \\ + \beta_{mp}^{RMW} \times RMW_t + \beta_{mp}^{CMA} \times CMA_t + \beta_{mp}^{WML} \times WML_t + \epsilon_{mpt} \end{aligned} \tag{1}$$

For each portfolio recommended by LLM m for investor profile p, we regress portfolio excess returns ($r_{mpt} - r_{ft}$) in month t on the returns to six well-known asset pricing factor portfolios: the market portfolio ($R_{mkt,t} - r_{ft}$), the small-minus-big (SMB) size portfolio, the high-minus-low (HML) value portfolio, the robust-minus-weak (RMW) operating profitability portfolio, the conservative-minus-aggressive (CMA) investment portfolio, and the winners-minus-losers (WML) momentum portfolio. We use the factors for developed markets from the website of Kenneth French.[21] For each portfolio, we obtain a measure of risk-adjusted returns (α_{mp}) and a measure of market risk (β_{mp}^{mkt}). We further compute idiosyncratic volatility (i.e., the risk component not accounted for by the six risk factors) as the standard deviation of the error term $\sigma(\epsilon_{mp})$.

3.5.3 Exposure

To assess the fit between portfolio composition and investor characteristics, we construct measures of exposure to asset classes, markets, and risk factors. Specifically, we obtain data on asset class and country exposure on an asset level and aggregate them to the portfolio level by applying weighted averages.[22] We distinguish between equity, fixed income, alternative assets, and cash as asset classes. Alternative assets include commodities, private equity, private debt, cryptocurrencies, and real estate (including REITs). We further aggregate country exposures to the MSCI market definitions (developed and emerging markets).[23] To test for potential home bias in the portfolio recommendations, we further obtain the portfolio weights of the investor profiles' home countries. Finally, we compute three measures capturing distinct portfolio risk components from the monthly portfolio time series described in the previous section. We measure total portfolio risk using the volatility of monthly portfolio returns. We measure market-wide and idiosyncratic risk using the market beta and idiosyncratic volatility derived from the six-factor model (equation 1).

[21] As a robustness test, we re-run the factor model specifications using region-specific factor portfolios; i.e., developed market factor portfolios for German and US profiles and emerging market factor portfolios for Chinese profiles. All portfolio data are obtained from Kenneth French.

[22] Exposure data are obtained from Bloomberg, Thomson Reuters Refinitiv Eikon, the Financial Modeling Prep (FMP) API, and through desk research (in that order).

[23] MSCI classifications can be found here: https://www.msci.com/our-solutions/indexes/market-classification. In addition to the MSCI classification, developed markets include Luxembourg and Liechtenstein. We do not investigate exposure to frontier and standalone markets as they feature marginally in the portfolios.

3.5.4 Performance

To assess the portfolios' historical risk-return profiles and performance, we compute three measures capturing the risk-adjusted performance of the portfolios, and one measure capturing the operating costs of implementing the portfolio. First, we compute average monthly excess returns. Second, we compute the Sharpe ratio as a simple risk-adjusted return measure. Third, we use the α_{mp} coefficients from the six-factor regression described in equation 1. Fourth, to account for administrative and management fees, we aggregate product-level total expense ratios (TER) to the portfolio level.

4 Results

Table 2 reports summary statistics for the various dependent variables we consider, grouped by the three suitability dimensions implementability, exposure, and performance. In this section, we investigate the hypotheses derived for each of the three suitability dimensions. For the baseline regressions, we use the following specifications:

$$DV_{mp} = \alpha + \beta X_m + \gamma X_p + u_{mp} \tag{2}$$

where DV_{mp} are the various dependent variables capturing different suitability aspects of a portfolio suggested by LLM m for investor profile p. X_m is a vector containing LLM characteristics (LLM type, license type, and size), and X_p is a vector containing profile characteristics (age, risk tolerance, home country, sustainability preferences). Tests of hypothesis 1 (general ability of LLMs to generate financial advice) will generally be based on summary statistics and γ coefficients, while tests of hypotheses 2 and 3 (relationship of specific LLM features with ability to generate financial advice) will generally be based on the vector of β coefficients.

Table A7 in the Appendix reports pairwise correlation coefficients for the main dependent variables. Some observations are worth noting. First, the implementability measures are correlated, which implies that they measure the same latent concept. Second, a higher exposure to developed markets and US securities is associated with higher data availability and fewer response errors. The correlation could suggest that LLMs more frequently misreport tickers for emerging market securities, which is why data availability is lower. Alternatively, this pattern could be a consequence of the dominance of the US and other developed markets in international capital markets and text corpora. Finally, there seems to be a relationship between the implementability measures and performance. Portfolios with more assets and higher data availability display higher risk-adjusted returns, while portfolios with more response errors display lower risk-adjusted returns (Sharpe ratios). Thus, we will account for differences in implementability when we investigate performance.

[Table 2 approximately here]

4.1 Implementability

In this section, we investigate whether LLMs are generally able to provide portfolio recommendations that can be implemented and whether LLM type and size are related to implementability.

Panel A of Table 2 reports summary statistics for the implementability measures. The number of recommended assets is generally moderately low (mean = 4.8, median = 5.0), but is at times as high as 73. Data is available for 93% of tickers on average (conditional on correctly specified tickers). We observe some kind of response error in 47% of the portfolios. While this number may seem high, we argue that implementability is still given for two reasons.

First, most errors can be considered minor in terms of their impact on investors' ability to implement the recommended portfolio. Of the erroneous suggestions, responses most commonly specified weights that did not add up to 100% (34%).[24] The second most frequent error was the lack of a specific recommendation (17%), which occurred either because the LLMs generated a generic response (mostly suggesting a low-cost passive strategy) or because no response was received for a prompt through the API. We find that two LLMs failed to generate specific recommendations for all 64 profiles, while 13 LLMs always provided specific recommendations and the remaining LLMs were able to generate specific recommendations for most profiles. In 7% of portfolios, no portfolio weights were specified for at least one of the securities. Finally, in 5% of portfolios, at least one of the suggested securities' tickers was wrongly specified, and in 4% of portfolios, no ticker was provided for at least one of the suggested securities.

Second, Tables A8 and A9 in the Appendix suggest that explicitly addressing the potential errors in the prompt significantly reduces the incidence of errors. We use a simple, zero-shot prompt in our main analyses to obtain results that are comparable across LLMs and can be easily implemented by retail investors. Related studies document that performance improvements can be achieved through prompt engineering (Lopez-Lira, 2024; Lopez-Lira and Tang, 2023). Thus, we assess the extent to which more sophisticated prompts reduce erroneous responses in our context. To do this, we use the original prompt and add a specific instruction to avoid the original response errors.[25] We then compare the error incidences of the original prompt to those of the adjusted prompts. Tables A8 (OLS) and A9 (logistical regression) in the Appendix report the resulting coefficients. The results suggest that using explicit correction prompts reduces the likelihood of errors by up to 21 percentage points or roughly half of the baseline incidence. Since the most common response errors do not jeopardize implementation

[24] We allow for a 10% tolerance, i.e., the indicator variable only takes on the value 1 if the sum of portfolio weights was larger than 110% or less than 90%.

[25] Of the 32 original LLMs, 19 LLMs were still available at the time of the second data collection. For each of the 1,064 recommendations (19 LLMs × 56 profiles), we obtain data for 6 different prompts (1 original and 5 correction prompts). Data were collected from August 22 to 27, 2024. As an example, the correction prompt for error type 1 (no portfolio weight specified) consists of the original prompt and the additional sentence "Please make sure to specify the portfolio weights for each recommended financial product."

and can be significantly reduced through simple correction requests, we conclude that most LLMs are generally able to provide financial advice that is in fact implementable.

Result H1.1. *LLMs are generally able to provide implementable financial advice.*

To assess how different LLM and profile characteristics correlate with the implementability of the LLMs' financial advice, we perform multivariate regressions of the implementability measures on LLM and profile characteristics (as specified in equation 2). As dependent variables, we use the number of assets, the degree of data availability for portfolio assets, and the incidence of response errors.

Table 3 reports the results of the regressions. To test hypothesis 2, we investigate differences in implementability between foundation models and fine-tuned models. The results provide support for our hypothesis that foundation models are better suited to generate financial advice that is more easily implementable. Specifically, we find that the number of portfolio assets suggested is higher and the incidence of response errors is lower for foundation versus fine-tuned models ($p < 0.01$, respectively). The slightly higher number of assets in foundation models versus fine-tuned models (4.8 vs. 4.3 assets) should not affect implementability.[26] The difference in error incidence between foundation and fine-tuned models is economically meaningful: the OLS coefficients suggest a difference of 7 percentage points. The corresponding logit specification suggests a decrease in the error odds ratio of 27% for foundation versus fine-tuned models (see Table A10 in the Appendix).[27] We find no significant difference in data availability conditional on correctly specified tickers.

To avoid our results being driven by outlier LLMs, we re-estimate the regressions with a restricted sample excluding LLMs with either 100% response errors or 0% data availability (3 LLMs; see Table A10 in the Appendix). The results remain qualitatively unchanged.[28] Thus, regarding hypothesis 2, we conclude that foundation models generate financial advice that is more easily implemented than fine-tuned models.

Result H2.1. *Foundation models recommend portfolios that are more easily implemented than fine-tuned models.*

To test hypothesis 3, we investigate differences between LLMs of different sizes on the three implementability dimensions. The results suggest that larger LLMs generate portfolio suggestions that are more easily implemented than smaller LLMs. While the number of portfolio assets increases (0.6 assets, $p < 0.01$), the average number of assets does not exceed five, which can be considered manageable (cf. Foerster et al., 2017). The likelihood of response

[26] Foerster et al. (2017) report an average of 5.2 funds in their Canadian retail investor sample (Table I).

[27] $e^{-0.315} - 1 = -0.270$

[28] We also apply a specification using a continuous size variable instead of a size dummy (see Table A11 in the Appendix). Using this specification, we find no significant difference in implementability between foundation models and fine-tuned models, which is not surprising given the strong collinearity between LLM type and size (all LLMs with more than 70B parameters are foundation models; see Table 1).

errors is negatively related to LLM size. Specifically, the incidence of response errors is 13 percentage points lower for larger LLMs. The corresponding logit specification suggests a difference of 44% in the error odds ratio between large and small LLMs (see Table A10 in the Appendix).[29] The results are replicated when excluding LLMs with either 100% response errors or 0% data availability (see Table A10 in the Appendix) or when using a continuous size variable instead of a size dummy in the regressions (see Table A11 in the Appendix).

Result H3.1. *Larger LLMs recommend portfolios that are more easily implemented than smaller LLMs.*

The coefficients on the investor profile characteristics are mostly in line with expectations. Specifically, LLMs recommend more assets for more risk-tolerant profiles, potentially because of the larger number of specialized equity funds as compared to bond funds (portfolios suggested to risk-tolerant profiles tilt more towards equity; see Table 5). Besides, suggestions for Chinese and German investor profiles display lower data availability and a higher likelihood of response errors, which is likely due to the higher prevalence of US securities in global capital markets and in the training data. Specifically, we prompt ChatGPT-4o, which we do not use in our main analysis, for the number of available securities by domicile country (China, Germany, and the US) and instrument type (equity ETFs, sustainable equity ETFs, fixed-income ETFs, and stocks). The results suggest that the recognized number of US-based ETFs exceeds those domiciled in China or Germany for all ETF types and by a wide margin (see Table A12 in the Appendix). Finally, portfolios recommended to sustainability-oriented profiles are more susceptible to response errors. This might reflect the fact that sustainability-oriented investment products are a relatively new phenomenon and hence do not feature as prominently in the training data as conventional investment products. The results are robust to the exclusion of outlier LLMs (see Table A10 in the Appendix).

[Table 3 approximately here]

4.2 Exposure

In this section, we investigate whether the LLMs' portfolio recommendations are generally in line with investor characteristics such as risk tolerance, risk capacity, and sustainability preferences (ESMA, 2018; SEC, 2019). We further study whether LLM type and size are related to the degree to which the recommendations take into account investor characteristics. While there is no generally accepted optimal solution for a portfolio's exposure, we assess whether portfolio exposures are suitable for the respective investor profiles in three ways. First, we compare the portfolio exposures as well as their sensitivity to investor characteristics to those of the benchmark robo-advisory portfolios. Second, we compare them to the exposures and sensitivities in the portfolios of retail investors without advice or with human financial ad-

[29] $e^{-0.578} - 1 = -0.439$

visors as reported in related studies (Bhattacharya et al., 2024, 2012; Foerster et al., 2017; Jacobs et al., 2014). Third, we assess whether the relationship between investor characteristics and exposure measures coincides with the principles of modern portfolio theory (Lintner, 1965).

Panel B of Table 2 reports summary statistics for the exposure variables.[30] The average equity share is 67%, while the average fixed income allocation is 30%. Alternative assets and cash only feature marginally in the average portfolio. Individual stocks make up 2% of portfolio values on average. This is broadly in line with the average exposures in robo-advisory portfolios (59% equity, 38% fixed income; see Table 4). The exposures of the LLM-generated recommendations are also consistent with figures reported in an audit study involving German advisors (Bhattacharya et al., 2012, p. 983), a study involving 175,000 clients of 5,000 Canadian advisory firms (Foerster et al., 2017, p. 10), and a study using data from a large US robo-advisor (Rossi and Utkus, 2024). On average, the Markowitz-type optimization algorithm used by Bhattacharya et al. (2012) recommended a slightly higher equity share (73%), lower fixed income share (6%), and substantially higher individual stock share (53%) than the LLMs in our sample. Foerster et al. (2017) report an average risky share of 68% for moderately risk-averse investors and 74% for the median investor advised by a human financial advisor.[31] Rossi and Utkus (2024) report an average equity share of 59% for their sample of relatively old robo-advisory clients (average age is 64; Rossi and Utkus, 2024, Table 2). The asset class breakdown is also broadly consistent with Jacobs et al. (2014), who consult previous literature and derive a consensus allocation of approximately 60% equity and 40% fixed income.

[Table 4 approximately here]

With respect to markets, the average portfolio allocates 7% to Chinese, 4% to German, and 65% to US securities. Correspondingly, the average share invested in the investor's home country is 49% for all profiles. The average portfolio allocates 88% to developed market securities and 12% to emerging market securities. The exposure to domestic equity is higher in LLM recommendations than in robo-advisory recommendations (43% vs. 31%).[32]

To assess whether portfolio exposure is suitable for a specific investor profile, we perform multivariate regressions of various exposure variables on LLM and profile characteristics (as specified in equation 2). As dependent variables, we use the share of equity and fixed income

[30] Table A13 in the Appendix reports summary statistics for the relevant variables for alternative factor model specifications. Specifically, we additionally employ region-specific factor portfolios and assume annual rebalancing.

[31] The risky share ranges from 37% for the least risk-averse clients to 75% for the most risk-averse clients.

[32] We investigate a potential home bias separately in section 4.4.1.

securities as measures of asset class exposure.[33] We use exposure to developed markets, as well as US and domestic securities as measures of geographical exposure. Finally, we measure exposure to risk using total risk (monthly volatility), market-wide risk (FF6 market beta), and portfolio-specific risk (idiosyncratic volatility). Table 5 reports the coefficients of the baseline specification.

[Table 5 approximately here]

Across LLMs, investors with high risk tolerance were recommended portfolios with a 30 percentage-point higher equity share ($p < 0.01$) than low-risk-tolerance investors, which translates to significantly higher portfolio risk (using all three risk measures). This difference is generally consistent with the benchmark robo-advisory portfolios and the results of related studies: While robo-advisors seem to distinguish more strongly between investors with high and low risk tolerance than LLMs (42 percentage points; see Table A14 in the Appendix), the algorithm used by Bhattacharya et al. (2012) recommends a slightly lower 25 percentage-point difference in equity shares for the most- versus least-risk-tolerant investors.[34] Foerster et al. (2017) document a 38 percentage-point difference in equity share for the most versus least risk tolerant clients of Canadian advisors.[35]

We further find that portfolios recommended to high-risk-tolerance investors tilt more towards emerging markets (4 percentage-point lower developed market exposure, $p < 0.01$). This pattern is consistent with higher tolerance for greater risk in emerging markets, for example due to political uncertainty or capital flight risk. Finally, we find a slightly higher prevalence of single stocks (3 percentage points, $p < 0.01$) among portfolios recommended to more risk-tolerant investors.

Consistent with a shift towards lower-risk securities closer to retirement, we also find that equity exposure decreases with age (10 percentage-point difference for 60-year-olds vs. 30-year-olds, $p < 0.01$) and fixed-income exposure increases with age (8 percentage-point difference, $p < 0.01$), which translates to lower portfolio risk. Given the high baseline equity shares, a difference of 10 percentage points between 60-year-olds and 30-years-olds strikes us as low (cf. Foerster et al., 2017). Corroborating this notion, the average difference in equity shares between 60-year-olds and 30-year-olds is 18 percentage points in the robo-advisory benchmark portfolios (see Table A14 in the Appendix). Notably, we find that the oldest investors are recommended a higher share of individual stocks by LLMs. In many cases,

[33] We do not use the exposure to alternative assets, mixed-asset securities or cash due to their marginal average shares (see Table 2).

[34] 63% recommended to high-risk-tolerance investors, 38% recommended to low-risk-tolerance investors (Bhattacharya et al., 2012, Table 2).

[35] See Table II, Panel A (Foerster et al., 2017).

this is due to LLMs recommending stocks known to pay high and consistent dividends as retirement approaches.[36]

To test for the relative importance of the different investor characteristics in determining portfolio risk, we run relative weight analyses on the central exposure measures. The weights reflect the relative contribution of each LLM and investor characteristic to the total predicted variance in the exposure variables (cf. Blaseg and Hornuf, 2024). For robustness, we further compute 95% confidence intervals for each weight based on 10,000 bootstrap samples. Table A17 in the Appendix reports raw and standardized weights of the various independent variables in regressions on the equity share, volatility, market risk, and idiosyncratic risk. If LLMs' recommendations are consistent with modern portfolio theory (Lintner, 1965), risk tolerance should be the most important contributor to explaining the variation in portfolio exposure. The results support this notion: Risk tolerance accounts for 84% of the explained variation in the equity share, 92% of the explained variation in monthly volatility, 88% of the explained variation in market risk, and 30% of the explained variation in idiosyncratic volatility. It dominates any other investor characteristic in the first three specifications. Only in idiosyncratic volatility regressions is risk tolerance dominated by the Chinese investor dummy. This pattern persists when we use region-specific factor portfolios (see Table A18 in the Appendix), which suggests that the portfolios recommended to Chinese investors are systematically related to risk factors other than the developed and emerging market risk factors commonly used in asset pricing studies.[37] A potential explanation is that prices of Chinese stocks are highly sensitive to the announcements of national regulation (The Economist, 2024a), thus causing them to be out of sync with general emerging market trends. Consistent with the LLM recommendations, risk tolerance is the dominant investor characteristic in portfolios recommended by robo-advisors both in our sample (see Table 6) and in a related study (Scherer and Lehner, 2023). However, in line with the regression coefficients, age seems to play a greater role in explaining portfolio exposure in robo-advisory portfolios than in LLM-generated portfolios.

Next, we investigate whether LLM-generated portfolio recommendations are "one-size-fits-all" solutions. Foerster et al. (2017) find that the portfolios recommended by Canadian advisors are driven by advisor fixed effects to a much greater extent than observed investor characteristics. Specifically, they reveal that the client's observable characteristics jointly explain only 12% of the cross-sectional variation in risky share, while advisor characteristics explain one-and-a-half times as much of the variation in the risky share. Table 6 displays relative weights for regressions of exposure variables on investor characteristics and LLM or robo-advisor fixed effects. For a sensible comparison, we only include portfolios for the 24

[36] The results reported in Table 5 are robust to the inclusion of implementability measures as covariates (see Table A15 in the Appendix) and to using region-specific factor portfolios and an annual rebalancing frequency (see Table A16 in the Appendix).

[37] The results are also robust to using a continuous size variable instead of size dummy (see Table A19 in the Appendix).

US and German investor profiles for which we obtain recommendations from robo-advisors. The standardized weights suggest that LLM fixed effects account for 21% of the explained variation in equity shares, 31% of the explained variation in volatility, 25% of the explained variation in market risk, and 53% of the explained variation in idiosyncratic volatility. Thus, investor characteristics account for a larger share of the explained variation in all but one specification, which suggests that LLMs provide more individual portfolio recommendations than the Canadian advisors in Foerster et al. (2017). The same holds true when the weights are compared to those from the robo-advisory portfolios (see Table 6). In our study, robo-advisor fixed effects account for 16% of the explained variation in equity share, 60% of the explained variation in volatility, 64% of the explained variation in market risk, and 94% of the explained variation in idiosyncratic volatility.

[Table 6 approximately here]

Taken together, the relation between the exposure variables and investor characteristics suggest that LLMs are generally capable of providing financial advice that takes into account investors' specific circumstances. All correlations are in line with principles of sound financial advice and are generally consistent with the recommendations of the robo-advisors in our sample, a rule-based Markowitz optimizer (Bhattacharya et al., 2012), professional Canadian advisors (Foerster et al., 2017), and the large robo-advisor studied by Rossi and Utkus (2024). In fact, investor characteristics account for more of the explained variation in the exposure measures in LLM-generated portfolios than in portfolios managed by the human advisors studied by Foerster et al. (2017) and the robo-advisors in our sample.

Result H1.2. *LLMs generally recommend portfolios that take into account relevant investor characteristics.*

In addition to the baseline regression, we assess the degree to which different LLM features affect the sensitivity of our exposure measures to the individual circumstances of the investor profiles. Because we show that risk tolerance is the main determinant of portfolio allocations both in theory and in our sample, we focus our analysis on the sensitivity of exposure to risk tolerance. To test hypotheses 2 and 3, we adjust the baseline regression specification (equation 2) by including interaction terms between the risk-tolerance dummy and (i) the foundation model dummy and (ii) the size dummy (Table 7 reports the results of these sensitivity analyses):

$$DV_{mp} = \alpha + \theta \, \mathbb{1}(\text{Risk tolerance} = \text{high}) \times \mathbb{1}(\text{Type} = \text{foundation model}) + \beta X_m + \gamma X_p + u_{mp} \tag{3}$$

$$DV_{mp} = \alpha + \theta \, \mathbb{1}(\text{Risk tolerance} = \text{high}) \times \mathbb{1}(\text{Size} > 60\text{B}) + \beta X_m + \gamma X_p + u_{mp} \tag{4}$$

The θ coefficients represent the additional difference in the exposure variables for investors with high and low risk tolerance. Positive coefficients suggest that foundation models and larger LLMs are more sensitive to an investor's risk tolerance than fine-tuned models and smaller LLMs.

Our results suggest that the recommendations provided by foundation models are more sensitive to an investor's risk tolerance than those provided by fine-tuned models (see Table 7). While the difference in equity shares between high- and low-risk-tolerance investor profiles amounts to 26 percentage points for fine-tuned models ($p < 0.01$), it amounts to 33 percentage points for foundation models. The difference between the LLM types of 7 percentage points is statistically significant at the 1% level. The higher sensitivity is also reflected in higher total, market-wide, and idiosyncratic risk.[38] Using the difference in equity shares observed in the robo-advisory portfolios (42 percentage points; see Table A14 in the Appendix) as a benchmark, the greater sensitivity in foundation models is considered more suitable in our context.

Result H2.2. *Foundation models recommend portfolios that are more suitable for the respective investor profiles than those of fine-tuned models.*

Regarding LLM size, our results suggest that larger LLMs are more sensitive to differences in investors' risk preferences: While the difference in equity shares between high- and low-risk-tolerance investors is 26 percentage points for small LLMs, it increases to 34 percentage points for large LLMs (increase of 8 percentage points, $p < 0.01$). Thus, while larger LLMs tilt more towards fixed income for all profiles (see Table 5), they differentiate to a greater degree with respect to an investor's risk tolerance. The larger difference in equity shares translates to a larger difference in volatility and market risk ($p < 0.1$, $p < 0.01$), but not idiosyncratic volatility.[39] Given these findings, we conclude that larger models recommend portfolios that are more sensitive to the investors' individual circumstances than smaller models, which more closely resembles exposures in the benchmark robo-advisor portfolios.

Result H3.2. *Larger LLMs recommend portfolios that are more suitable for the respective investor profiles than smaller LLMs.*

[Table 7 approximately here]

4.3 Performance

In this section, we investigate the historical performance of the LLMs' portfolio recommendations. Because the performance evaluations are by necessity based on backtesting, it is

[38] Both results are robust to using region-specific factor models and assuming annual rebalancing (see Table A20 in the Appendix), as well as using a continuous size variable (see Table A21 in the Appendix).

[39] The results are robust to using region-specific factor models and assuming annual rebalancing (see Table A20 in the Appendix), as well as using a continuous size variable (see Table A21 in the Appendix).

conceivable that LLMs base their portfolio allocation decisions on previous returns. To account for such look-ahead bias, we include an LLM's information cut-off date in our regression analyses. In addition, we run an experiment around an update to GPT-4 to assess whether it adjusts its portfolio recommendations to recent price developments.

Panel C of Table 2 shows summary statistics for the performance measures. The average monthly excess return is 0.36%, the average annual Sharpe ratio is 0.35, six-factor alphas are slightly negative on average (-0.09%) and range from -0.9% to 3.1%.[40] The average LLM-recommended portfolio underperformed the overall market, which experienced average monthly excess returns of 0.79% and a Sharpe ratio of 0.62 over the same horizon (2010 to 2023). The average performance figures remain qualitatively unchanged when we employ region-specific factor portfolios or assume annual rebalancing (see Table A13 in the Appendix).

Looking only at German and US profiles (Table 4), LLM-recommended portfolios display higher average excess returns (0.35% vs. 0.19%, $p < 0.01$) and Sharpe ratios (0.35 vs. 0.24, $p < 0.01$), but lower six-factor alphas (-0.12% vs. -0.08%, $p < 0.01$) than robo-advisors. Table 8 reports coefficients for regressions of historical portfolio performance on an indicator variable for LLM-generated recommendations and investor characteristics as control variables. The table reports the regression coefficients for the full sample, as well as different sub-samples. Table A22 in the Appendix reports the full regression outputs and Figure 3 displays the distributions of the three performance measures. In line with the descriptive results, LLM portfolios display higher excess returns (16 basis points, $p < 0.01$) and Sharpe ratios (0.11, $p < 0.01$), but lower six-factor alphas (5 basis points, $p < 0.01$). Thus, LLM-generated portfolios derive their superior performance from commonly priced risk factors to a greater extent than robo-advisory portfolios. The coefficients are highly robust for the various sub-samples. Notably, LLM portfolios do relatively better for older investors, US investors, and investor profiles with sustainability preferences.

[Table 8 and Figure 3 approximately here]

Taken together, the performance measures generally support the notion that LLMs recommend portfolios whose performance is no worse than that of professional robo-advisory portfolios.

Result H1.3. *LLMs recommend portfolios whose historical risk-adjusted performance is on par with professionally managed benchmark portfolios.*

To assess how different LLM and profile characteristics correlate with the performance measures, we perform multivariate regressions of the performance variables on LLM and profile characteristics (according to equation 2). To address concerns over look-ahead bias, we additionally account for the recency of an LLM's training data. We consider a dummy

[40] Point estimates for the six-factor alpha and market beta measures are replaced with zeros if they are not statistically significant at the 10% level.

variable indicating whether an LLM's training data cutoff falls within 6 months from the time of our initial analysis (April 2024) or not.[41]

Table 9 reports the results of the main regression specifications. Regarding hypothesis 2, we find that foundation models generate portfolios with slightly lower Sharpe ratios ($p < 0.1$) when information cutoff is considered, but no difference in any of the other performance measures. We run two robustness tests. First, we re-estimate all specifications to include the data availability and response error variables from the implementability issues. The concern is that (a) securities for which there is no historical price data and (b) securities that are associated with a response error (e.g., because of a mis-specified ticker) are systematically different in terms of performance. Table A23 in the Appendix reports the resulting coefficients. Using this specification, there is no longer any difference in performance between foundation and fine-tuned models. Second, we use region-specific factor portfolios and allow for annual rebalancing (see Table A24 in the Appendix). Using this specification, we do not measure any significant difference in performance between foundation and fine-tuned models. Hence, we find no evidence of superior performance in foundation models as compared to fine-tuned models. If anything, risk-adjusted performance as measured by annual Sharpe ratios is slightly lower in portfolios recommended by foundation models in one specification.

Result H2.3. *We find no evidence that foundation models recommend portfolios with superior risk-adjusted returns than fine-tuned models.*

Regarding hypothesis 3, we find that larger LLMs recommend portfolios with significantly lower Sharpe ratios ($p < 0.01$) when information cutoff is considered, but no difference in monthly excess returns and FF6 alphas.[42] When taking into account data availability and response errors (Table A23 in the Appendix), the difference in Sharpe ratios persists ($p < 0.05$). Hence, we find no evidence of superior performance of larger LLMs. If anything, risk-adjusted performance as measured by annual Sharpe ratios is lower in portfolios recommended by larger LLMs.

Result H3.3. *We find no evidence that larger LLMs recommend portfolios with superior risk-adjusted returns than smaller LLMs.*

[Table 9 approximately here]

The results reported in Table 9 suggest that LLMs with a more recent information cutoff generate lower Sharpe ratios ($p < 0.01$) than LLMs with no exposure to recent information. While we find no significant difference in six-factor alphas in most specifications, six-factor alphas are slightly higher (3 basis points, $p < 0.05$) when more recent information is available

[41] Some developers do not provide information on the information cutoff dates of their LLMs. In these cases, we use estimates provided in online forums as a rough indication of the information cutoff.

[42] The results remain unchanged when we use region-specific factor portfolios (see Table A24 in the Appendix).

only when we use region-specific factor portfolios and assume annual rebalancing (see Table A24 in the Appendix). Thus, our results imply that access to more recent information does not generally improve historical performance. If anything, performance as measured through simpler measures such as the Sharpe ratio is worse for LLMs with more recent access to information. These results are somewhat surprising; one would expect look-ahead bias to materialize more in less-complex performance measures such as excess returns or Sharpe ratios. While a recent experiment confirms that advanced LLMs are able to compute CAPM betas from historical data if prompted (Korinek, 2023), look-ahead bias is most likely driven by news coverage of particularly well-performing stocks. Consistent with this notion, among the most frequently recommended individual stocks—besides well-known American tech stocks—are those with good recent performance and high retail investor attention such as Netflix and Nvidia.[43]

As a more explicit test of whether LLMs take into account the recent performance of securities when making investment recommendations, we exploit an information update to GPT-4, one of the most prominent and publicized LLMs. We initially collected portfolio recommendations from GPT-4 using very similar prompts for a subset of profiles in May 2023. At that time, the LLM's information cutoff was September 2021. After OpenAI released an updated version of GPT-4 drawing on data up until January 2022, we collected the portfolio recommendations for the same profiles and using the same prompts again in October 2023. This allows us to directly compare the performance of the portfolios recommended before and after the information update. If GPT-4 explicitly takes into account past performance, portfolios suggested by the updated GPT-4 version should outperform the initial recommendations' performance. This should particularly hold for the period between October 2021 and January 2022, for which the updated recommendations (but not the initial ones) could have relied on past performance. Figure A1 in the Appendix displays the performance of the GPT-suggested portfolios both for the initial recommendation (based on data up until September 2021) and the updated recommendation (based on data up until January 2022). It does not appear that the updated portfolio recommendations outperform the initial recommendations, neither over the entire sample period (September 2021 to May 2023), nor over the period from September 2021 to January 2022. Taken together, these two analyses alleviate the concern that our performance analyses are skewed due to look-ahead bias.

4.4 Biases

The previous sections demonstrate that LLMs are generally capable of providing financial advice similar to that a human advisor. The question we pose in this section is whether LLMs also display biases that have been documented in their human counterparts. LLMs are trained on vast amounts of text produced by humans. If the training data, which includes text

[43] The top 10 stocks include Apple (recommended 63 times), Amazon (51), Tesla (44), Google (35), Microsoft (34), Berkshire Hathaway (22), Meta (14), Netflix (12), Nvidia (10), and LendingTree (6).

from diverse sources such as books, chat forums, social media platforms, and news articles (Zhao et al., 2023), contains biases, it is conceivable that LLMs inherit some of the biases contained in the training data (Kordzadeh and Ghasemaghaei, 2022; The Economist, 2024d).

We focus on two distinct biases that could occur in the context of LLM-generated financial advice. First, we investigate home bias; i.e., an excessive allocation to domestic securities, which is a common feature of both institutional and retail investors' portfolios (Ardalan, 2019). Second, we study whether LLMs inherit gender-based discrimination. A related study shows that social media algorithms inadvertently target male users when displaying a gender-neutral ad for STEM training and jobs (Lambrecht and Tucker, 2019). Similarly, AI-based human resources tools consistently favor men over women (The Economist, 2024e). In the financial advice context, studies document lower stock market participation and risk tolerance among women (Croson and Gneezy, 2009; Halko et al., 2012). In addition, a recent audit study documents that women receive portfolio recommendations skewed toward individual and local securities, likely as a consequence of statistical discrimination (Bhattacharya et al., 2024). Thus, we study whether—holding other investor characteristics constant—the investor's gender affects the quality of portfolio recommendations.

4.4.1 Home bias

As a measure for country-level home bias, we compute the gap between the share of securities d_i from the domestic country i in the portfolio and the share of securities m_i from country i in the global market portfolio (Cooper et al., 2018; Lau et al., 2010). We use country weights in the MSCI All-Country World Investable Markets Index as measures for m_i.[44] The index is a widely used market benchmark for equity ETFs. According to MSCI, the index provider, it covers 99% of the investable global equity market.

US securities make up 63% of the index, while Chinese and German securities make up 2% each.[45] In portfolio recommendations by LLMs, Chinese securities account for 29% of all assets and 23% of the equity portion of Chinese investors' portfolios, suggesting a weight-gap of 21 percentage points (see Panel A of Table 10). German securities account for 11% of all assets recommended by LLMs and 9% of the equity portion of German investors' portfolios, suggesting a weight-gap of 7 percentage points (see Panel A of Table 10). US securities account for 72% of all assets recommended by LLMs and 73% of the equity portion of US

[44] Weights as of July 15, 2024. https://www.msci.com/research-and-insights/visualizing-investment-data/acwi-imi-complete-geographic-breakdown

[45] As an alternative specification, we use World Bank data on global market capitalization figures, which do not take into account the investment grade of stocks (Lau et al., 2010). Thus, whereas Chinese securities feature more prominently (14%) and the German share remains unchanged (2%), US securities feature less prominently (49%). Our results are qualitatively unchanged when using this alternative benchmark specification.

investors' portfolios, suggesting a weight-gap of 11 percentage points.[46] All differences are statistically significant at the 1% level. The results remain qualitatively unchanged when we account for the potentially distorting effect of erroneous tickers (panel B) and different risk tolerance levels (panel C).

To compare the severity of the home bias for different countries, two alternative weight gap measures have been suggested in the literature (Cooper et al., 2018). First, we divide the gap $d_i - m_i$ by the baseline weight m_i to assess the relative deviation from the benchmark weight. Using this specification, the home bias is most severe for China, with the actual weight equal to 7 to 9 times the benchmark weight. Second, we divide the gap $d_i - m_i$ by the international portion of the market portfolio $1 - m_i$ to assess how much of the international allocation is consumed by domestic securities. Using this measure, the home bias is most severe for US portfolios, closely followed by Chinese portfolios. Either way, we document a substantial home bias in portfolios for all three home countries.[47]

[Table 10 approximately here]

To gauge the distribution of excessive domestic allocation, Figure A2 in the Appendix displays density plots for the domestic allocation and various sub-samples by home country. The vertical red lines depict the baseline weights m_i. The density plots confirm our findings from the various home bias measures: while there is home bias for all home countries, it is more pronounced for the Chinese and US profiles and less so for German profiles, for which the mode of the distribution is below the benchmark weight. This finding might result from LLMs recommending European rather than national stock and bond funds for domestic exposure in the case of German investors.

Table A26 in the Appendix presents equivalent analyses for portfolio recommendations obtained from the robo-advisors. The results suggest that robo-advisory recommendations do not suffer from home bias. For US investors, the domestic equity share is either not statistically significantly different from the benchmark weight m_i or even slightly lower (59% vs. 62%, $p < 0.01$). For German investors, the suggested domestic equity share (3%) marginally exceeds the benchmark weight of 2% ($p < 0.01$).

One potential explanation for the home bias is that capital cannot easily move between international borders (Bekaert and Wang, 2009). Among the three countries studied in our analyses, Germany and the US rank among the countries with the least restrictions on international capital movements, while China is among the most restrictive jurisdictions due to its extensive capital controls (according to the International Monetary Fund's capital account

[46] The results are virtually identical whether we use all US profiles (profiles 13 through 24 and profiles 49 through 64) or only the base profiles without reference to gender or investor experience (profiles 13 through 24).

[47] Table A25 in the Appendix displays coefficients from regressions of the home bias measures on LLM and profile characteristics. The results suggest that the recommendations of larger LLMs suffer more from home bias than those of smaller LLMs.

openness index; cf. IMF, 2016). Thus, if the portfolios recommended to Chinese investors include international securities that Chinese investors cannot easily invest in due to capital controls, the documented home bias in the recommended portfolios underestimates the true home bias for Chinese investors. Because the average international exposure among Chinese investors is 71% (median 93%) and the average US exposure is 50% (median 58%) in our sample, the true home bias for Chinese investors is likely even more pronounced.

While the home bias documented in the LLMs' recommendations is substantial, it is no worse than the home bias documented among human advisors: Bhattacharya et al. (2024) document that Hong Kong based auditor investors were exclusively recommended securities domiciled in Hong Kong in 39% of meetings with a human advisor. In a related study, an automated Markowitz optimizer tool recommended 30% domestic equity exposure to a sample of German retail investors (Mullainathan et al., 2012). Finally, Foerster et al. (2017) find that the median investor in their Canadian advisee sample allocates 60% of their portfolio to Canadian equities, which reflects extreme home bias (Canada's weight in the MSCI All-Country World Index is 3%). The recommendations generated by LLMs in our sample also represent an improvement over un-advised retail investor portfolios: the average German retail investor allocated 52% of their equity portion to German securities prior to the advice intervention (Bhattacharya et al., 2012, p. 983). Besides, we document in a supplementary analysis that explicitly referencing potential home bias in the LLM prompt significantly reduces domestic allocation (see Table A27 in the Appendix).[48]

4.4.2 Gender-based discrimination

Eight of the investor profiles we define explicitly vary the investor's gender (see Table A2 in the Appendix). The profiles also vary with respect to risk tolerance (high vs. low) and age (between 30 and 60). This section investigates whether LLMs systematically adjust their recommendations to an investor's gender, controlling for risk tolerance and age.

Given that risk tolerance and stock market participation differ by gender (Croson and Gneezy, 2009; Halko et al., 2012), we focus our analyses on measures capturing the equity share and risk level of the recommended portfolios. Specifically, we investigate differences in the equity share, market risk, and idiosyncratic volatility. In addition, we integrate findings from a related study on gender-based discrimination in financial advice (Bhattacharya et al., 2024) and investigate gender differences in the domestic portfolio weights and shares of individual stocks.[49]

[48] In addition to the standard prompt, we add the following sentence: "Please make sure the recommendations do not suffer from home bias, which refers to the tendency of investors to excessively allocate portfolios to domestic assets."

[49] While Bhattacharya et al. (2024) document substantial gender differences in the recommendations of Hong-Kong-based financial planning firms, d'Astous et al. (2024) find no general gender differences in the recommendations of Canadian planners. None of the robo-advisors in our sample differentiate by gender.

Table A28 in the Appendix reports averages of the exposure and risk measures by profile and gender, as well as the test statistics of non-parametric rank tests for differences between the male and female exposures. Taking into account risk tolerance and age, there is no significant difference in any of the measures by gender on average. However, the aggregate pattern masks some heterogeneity for different recommendations. Figure 4, which plots the distributions of the gender gap in equity shares, suggests that while there is no systematic gender gap, there are recommendations with a gender gap in equity shares of up to 50 percentage points in either direction. As Table A29 in the Appendix shows, we find no LLM-specific differences in exposure to markets or risk components in recommendations to male and female investors.

[Figure 4 approximately here]

5 Conclusion and outlook

This study explores the potential of LLMs to provide financial advice. We elicit 2,048 portfolio recommendations for 64 investor profiles from 32 different LLMs. To assess whether LLMs produce suitable portfolio recommendations, we investigate their implementability, the appropriateness (given investor characteristics) of the portfolios' exposures, and their historical risk-return profiles. Based on existing evidence on the capabilities of LLMs, we hypothesize that LLMs are generally capable of providing suitable financial advice. We further hypothesize that foundation models and larger LLMs are better suited to generate portfolio recommendations than fine-tuned and smaller LLMs.

Our results suggest that LLMs are capable of providing suitable financial advice. Most LLMs recommended portfolios that (i) can in fact be implemented, (ii) take into account the individual circumstances of the investor, and (iii) exhibit historical performance on par with that of professionally managed benchmark portfolios. Turning to heterogeneity with respect to LLM features, our results suggest that foundation models and larger LLMs generate financial advice that is more easily implemented and better suited to the investor's circumstances than fine-tuned models and smaller LLMs.[50] We do not observe any systematic performance differences between the various LLM features.

There are obvious risks associated with the unrestricted use of LLMs in financial advice provision, given that mistakes can cause real harm to retail investors. As an example, Knight Capital Group, a former financial services firm that specialized in executing stock trades, lost more than $400 million when it went live with faulty software in 2012 (Reuters, 2012). To mitigate the potential of faulty recommendations, providers might resort to limiting the range of potential products to be recommended, as well as the portfolio weights assigned to each one to avoid faulty recommendations. As of now, rule-based robo-advisors assign weights to a set of pre-selected investment products (Rühr et al., 2019). Narrowing the range of potential

[50] Table A30 in the Appendix provides a ranking of all 32 LLMs based on a simple weighting of the three suitability dimensions we study in this paper.

portfolio recommendations could be achieved through "knowledge injection" (Ovadia et al., 2023). One way of operationalizing this is to adjust the LLMs' weights to the financial advice context through fine-tuning. Fine-tuning models to the financial context is challenging because high-quality financial data is hard to obtain (Lo and Ross, 2024), and evidence on the performance of finance-specific LLMs is sparse and conflicting. On the one hand, BloombergGPT, an LLM trained in part on Bloomberg's proprietary data, performs better at finance-specific sentiment analysis than comparable LLMs (Wu et al., 2023). On the other hand, LLMs fine-tuned to the financial context (BloombergGPT, FinBERT) have been shown to underperform cutting-edge general-purpose LLMs such as GPT4 in analyzing financial text (Li et al., 2023). Another way of injecting information specific to the financial advice context is through in-context learning, mostly operationalized through retrieval-augmented generation (RAG). By adding context to each query, an LLM's responses can be steered without affecting the pre-trained weights (Ovadia et al., 2023). In practice, financial advisors could provide model portfolios for investors with different individual circumstances, which the LLM will then prioritize as input for the provision of financial advice.

Another challenge associated with the use of LLMs concerns ethical advice-giving. A crucial issue is LLMs' tendency toward sociopathic behavior. Above, we describe the capability to simulate a personal relationship as a potential benefit of employing LLMs for financial advice; however, this may also pose challenges (Lo and Ross, 2024). Because their communication style is independent of the actual quality of the recommendations, LLMs are able to communicate bad recommendations with the same affect as good ones, which potentially erodes trust in the long run. Another related issue is that of self-serving advice. For example, Agnew et al. (2018) show that confirming a client's views (however biased they may be) increases advice-following. For human advisors, it has been shown that nudges such as requiring advisors to take a banker's oath analogous to the Hippocratic oath in medicine can reduce self-serving advice-giving (Weitzel and Kirchler, 2023). Future work is still needed to investigate ways of ensuring ethical behavior in LLMs in the financial advice context.

Some of the results in this paper should be interpreted in light of the limitations of our approach. Most notably, our performance analyses are based on historical data from 2010 to 2023. We employ this approach because we are interested in the long-term performance of the suggested portfolios. However, because LLMs are trained on large corpuses of texts mostly drawn from the internet, it is conceivable that portfolio recommendations generated by these LLMs take into account the previous price evolution of specific investment products. If that were the case, then our performance measures may overestimate the true potential of the recommendations due to look-ahead bias. While a true long-term out-of-sample performance test can only be conducted in a decade or two, two of our findings may help alleviate this concern. First, we find that recommendations generated by LLMs with more recent access to the internet do not outperform recommendations of LLMs with earlier information cutoffs when measured by simple risk-adjusted performance measures. Second, an experiment involving an update to GPT-4's information base in 2023 suggests that portfolios recommended after the information update do not outperform portfolios recommended before the information update. One explanation for this is that current LLMs are not sufficiently adept at conducting complex quantitative analyses (such as computing risk-adjusted returns on a large scale). Another explanation could be that LLMs place more emphasis on fundamental indicators than on recent returns when making portfolio recommendations (cf. Hornuf et al., 2025).

In light of our findings and the risks described above, it is unlikely that LLMs will be used without restrictions as financial advisors in the near future. Instead, they could be integrated as conversational agents into existing robo-advisors or assign investors one of several predefined model portfolios based on conversational risk profiling.

References

Aghion, P. and Bolton, P. (1992). An incomplete contracts approach to financial contracting. *The Review of Economic Studies*, 59(3):473–494.

Agnew, J. R., Bateman, H., Eckert, C., Iskhakov, F., Louviere, J., and Thorp, S. (2018). First impressions matter: An experimental investigation of online financial advice. *Management Science*, 64(1):288–307.

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

Ardalan, K. (2019). Equity home bias: A review essay. *Journal of Economic Surveys*, 33(3):949–967.

August, T., Chen, W., and Zhu, K. (2021). Competition among proprietary and open-source software firms: The role of licensing in strategic contribution. *Management Science*, 67(5):3041–3066.

Barron's (2024). Best Robo-Advisors. https://www.barrons.com/articles/best-robo-advisors-c2b901fe. [online; accessed 4 December 2024].

Beckmann, L., Beckmeyer, H., Filippou, I., Menze, S., and Zhou, G. (2024). Unusual Financial Communication-Evidence from ChatGPT, Earnings Calls, and the Stock Market. *Available at SSRN 4699231*.

Bekaert, G. and Wang, X. S. (2009). Home bias revisited. *Available at SSRN 1344880*.

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.

Bergstresser, D., Chalmers, J. M. R., and Tufano, P. (2009). Assessing the costs and benefits of brokers in the mutual fund industry. *Review of Financial Studies*, 22(10):4129–4156.

Bhattacharya, U., Hackethal, A., Kaesler, S., Loos, B., and Meyer, S. (2012). Is unbiased financial advice to retail investors sufficient? Answers from a large field study. *Review of Financial Studies*, 25(4):975–1032.

Bhattacharya, U., Kumar, A., Visaria, S., and Zhao, J. (2024). Do women receive worse financial advice? *The Journal of Finance*, 79(5):3261–3307.

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.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Bucher-Koenen, T. and Koenen, J. (2015). *Do seemingly smarter consumers get better advice?* MEA discussion paper.

Calvet, L. E., Campbell, J. Y., and Sodini, P. (2009). Fight or Flight? Portfolio Rebalancing by Individual Investors. *The Quarterly Journal of Economics*, 124(1):301–348.

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

Chalmers, J. and Reuter, J. (2020). Is conflicted investment advice better than no advice? *Journal of Financial Economics*, 138(2):366–387.

Christoffersen, S. E. K., Evans, R., and Musto, D. K. (2013). What do consumers' fund flows maximize? evidence from their brokers' incentives. *The Journal of Finance*, 68(1):201–235.

Cook, T. R., Kazinnik, S., Hansen, A. L., and McAdam, P. (2023). Evaluating local language models: An application to financial earnings calls. *Available at SSRN 4627143*.

Cooper, I. A., Sercu, P., and Vanpée, R. (2018). A measure of pure home bias. *Review of Finance*, 22(4):1469–1514.

Coval, J. D. and Moskowitz, T. J. (1999). Home bias at home: Local equity preference in domestic portfolios. *The Journal of Finance*, 54(6):2045–2073.

Cready, W. M. and Gurun, U. G. (2010). Aggregate market reaction to earnings announcements. *Journal of Accounting Research*, 48(2):289–334.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474.

D'Acunto, F., Prabhala, N., and Rossi, A. G. (2019). The promises and pitfalls of robo-advising. *Review of Financial Studies*, 32(5):1983–2020.

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.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., et al. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

d'Astous, P., Gemmo, I., and Michaud, P.-C. (2024). The quality of financial advice: What influences recommendations to clients? *Journal of Banking & Finance*, 169 (107291).

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].

Eulerich, M., Sanatizadeh, A., Vakilzadeh, H., and Wood, D. A. (2024). Is it all hype? chatgpt's performance and disruptive potential in the accounting and auditing industries. *Review of Accounting Studies*, 29(3):2318–2349.

ExtraETF (2023). Robo-Advisor Markt in Deutschland. https://extraetf.com/de/wissen/robo-advisor-markt-in-deutschland. [online; accessed 4 December 2024].

Fairhurst, D. J. and Greene, D. (2025). How Much Does ChatGPT Know About Finance? *Financial Analysts Journal*, 81(1):12–32.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2018). Choosing factors. *Journal of Financial Economics*, 128(2):234–252.

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

Foerster, S., Linnainmaa, J. T., Melzer, B. T., and Previtero, A. (2017). Retail financial advice: does one size fit all? *The Journal of Finance*, 72(4):1441–1482.

Forbes (2024). Top Robo-Advisors by Assets Under Management (AUM). https://www.forbes.com/advisor/investing/top-robo-advisors-by-aum/. [online; accessed 4 December 2024].

Frederickson, J. R. and Zolotoy, L. (2016). Competing earnings announcements: Which announcement do investors process first? *The Accounting Review*, 91(2):441–462.

French, K. R. and Poterba, J. M. (1991). Investor diversification and international equity markets. *American Economic Review (Papers and Proceedings)*, 81(2):222–226.

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.

Gennaioli, N., Shleifer, A., and Vishny, R. (2015). Money doctors. *The Journal of Finance*, 70(1):91–114.

Germann, M. and Merkle, C. (2023). Algorithm aversion in delegated investing. *Journal of Business Economics*, 93(9):1691–1727.

Germann, M., Mertes, L., Weber, M., and Loos, B. (2025). Trust and delegated investing: A money doctors experiment. *Review of Finance*, 29(1):75–102.

Halko, M.-L., Kaustia, M., and Alanko, E. (2012). The gender effect in risky asset holdings. *Journal of Economic Behavior & Organization*, 83(1):66–81.

Hansen, A. L. and Kazinnik, S. (2023). Can ChatGPT decipher Fedspeak? *Available at SSRN 4399406*.

Heinberg, A., Hung, A., Kapteyn, A., Lusardi, A., Samek, A. S., and Yoong, J. (2014). Five steps to planning success: Experimental evidence from US households. *Oxford Review of Economic Policy*, 30(4):697–724.

Helms, N., Hölscher, R., and Nelde, M. (2022). Automated investment management: Comparing the design and performance of international robo-managers. *European Financial Management*, 28:1028–1078.

Hens, T. and Nordlie, T. (2024). How good are llms in risk profiling? *Swiss Finance Institute Research Paper*, (24-30).

Hirshleifer, D., Lim, S. S., and Teoh, S. H. (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance*, 64(5):2289–2325.

Hornuf, L., Schmitt, M., and Stenzhorn, E. (2022). The local bias in equity crowdfunding: Behavioral anomaly or rational preference? *Journal of Economics & Management Strategy*, 31(3):693–733.

Hornuf, L., Streich, D. J., and Töllich, N. (2025). Domain-specific knowledge and llm performance: Evidence from a portfolio allocation experiment. *mimeo*.

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

IMF (2016). Capital Account Openness Index (April 2016). https://www.imf.org/external/datamapper/datasets/CL. [online; accessed 19 August 2024].

Ivković, Z. and Weisbenner, S. (2005). Local does as local is: Information content of the geography of individual investors' common stock investments. *The Journal of Finance*, 60(1):267–306.

Jacobs, H., Müller, S., and Weber, M. (2014). How should individual investors diversify? an empirical evaluation of alternative asset allocation policies. *Journal of Financial Markets*, 19:62–85.

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.

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

Kim, A., Muhn, M., and Nikolaev, V. (2023b). From transcripts to insights: Uncovering corporate risks using generative AI. *arXiv preprint arXiv:2310.17721*.

Kim, A., Muhn, M., and Nikolaev, V. V. (2024). Financial statement analysis with large language models. *Chicago Booth Research Paper Forthcoming, Fama-Miller Working Paper*.

Kordzadeh, N. and Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Lambrecht, A. and Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7):2966–2981.

Lau, S. T., Ng, L., and Zhang, B. (2010). The world price of home bias. *Journal of Financial Economics*, 97(2):191–217.

Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., and Shah, S. (2023). 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.

Linnainmaa, J. T., Melzer, B. T., and Previtero, A. (2021). The misguided beliefs of financial advisors. *The Journal of Finance*, 76(2):587–621.

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The Journal of Finance*, 20(4):587–615.

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.

Lo, A. W. and Foerster, S. R. (2023). In Pursuit of the Perfect Portfolio: Princeton University Press, 2021.

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. (2024). *The Predictive Edge: Outsmart the Market Using Generative AI and ChatGPT in Financial Forecasting*. John Wiley & Sons.

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

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

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, 7(1):77–91.

Mitchell, O. S. and Lusardi, A. (2022). Financial literacy and financial behavior at older ages. In *The Routledge Handbook of the Economics of Ageing*, pages 553–565. Routledge.

Morningstar (2023). Your Guide to Getting Started with Robo-Investing. https://www.morningstar.com/specials/your-guide-to-getting-started-with-robo-investing. [online; accessed 4 December 2024].

Mullainathan, S., Noeth, M., and Schoar, A. (2012). The market for financial advice: An audit study. In *NBER Working Papers 17929, National Bureau of Economic Research.*

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.

Ovadia, O., Brief, M., Mishaeli, M., and Elisha, O. (2023). Fine-tuning or retrieval? Comparing knowledge injection in LLMs. *arXiv preprint arXiv:2312.05934*.

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

Philippon, T. (2016). The fintech opportunity. Technical report, National Bureau of Economic Research.

Reuters (2012). Knight Capital posts $389.9 million loss on trading glitch. https://www.reuters.com/article/idUSBRE89G0HJ/. [online; accessed 19 August 2024].

Reuters (2023). Morgan Stanley to launch AI chatbot to woo wealthy. https://www.reuters.com/technology/morgan-stanley-launch-ai-chatbot-woo-wealthy-2023-09-07/. [online; accessed 19 August 2024].

Rossi, A. G. and Utkus, S. (2024). The diversification and welfare effects of robo-advising. *Journal of Financial Economics*, 157:103869.

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.

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., and Matarić, M. (2023). Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

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

Scherer, B. and Lehner, S. (2023). Trust me, I am a Robo-advisor. *Journal of Asset Management*, 24(2):85–96.

Scherer, B. and Lehner, S. (2025). What drives robo-advice? *Journal of Empirical Finance*, 80, 101574.

Schmidt, D. (2019). Distracted institutional investors. *Journal of Financial and Quantitative Analysis*, 54(6):2453–2491.

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. [online, accessed 12 November 2024].

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

Statistisches Bundesamt (Destatis) (2024). Vermögen und Schulden: Einkommen, Konsum und Lebensbedingungen. Retrieved December 17, 2024, from Statistisches Bundesamt.

Stolper, O. and Walter, A. (2019). Birds of a feather: The impact of homophily on the propensity to follow financial advice. *Review of Financial Studies*, 32(2):524–563.

Stolper, O. A. and Walter, A. (2017). Financial literacy, financial advice, and financial behavior. *Journal of Business Economics*, 87(5):581–643.

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

Tao, R., Su, C.-W., Xiao, Y., Dai, K., and Khalid, F. (2021). Robo advisors, algorithmic trading and investment management: wonders of fourth industrial revolution in financial markets. *Technological Forecasting and Social Change*, 163:120421.

The Economist (2024a). Buy, buy, buy ... sell! October 19, 2024.

The Economist (2024b). Death by LLM: The first casualties of generative AI offer lessons for other businesses. November 23, 2024.

The Economist (2024c). Freedom to tinker: AI models benefit from being open-source. November 9, 2024.

The Economist (2024d). Oh, the things AI can do. August 17, 2024.

The Economist (2024e). Risks and regulations — Artificial intelligence needs regulation. But what kind, and how much? August 24, 2024.

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

US Bureau of Labor Statistics (2024). Earnings and Wages: Current Population Survey (CPS). Retrieved December 17, 2024, from U.S. Bureau of Labor Statistics.

Weitzel, U. and Kirchler, M. (2023). The banker's oath and financial advice. *Journal of Banking & Finance*, 148:106750.

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*.

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*.

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: LLM overview

LLM	Developer	Model type	Foundation model	License type	Size (B)	Information cut-off
GPT-4-Turbo	OpenAI	Foundation		Proprietary	1000	Dec-23
Gemini-Pro	Google DeepMind	Foundation		Proprietary	600	Apr-23
Bard-Jan-24-Gemini-Pro	Google DeepMind	Foundation		Proprietary	540	Real-time updates
Jurassic-2 Ultra	AI21 labs	Foundation		Proprietary	178	Unknown
Mistral-Medium	Mistral AI	Foundation		Proprietary	175	Unknown
GPT-3.5-Turbo	OpenAI	Foundation		Proprietary	175	Sep-21
Claude 2.0	Anthropic	Foundation		Proprietary	130	Unknown
Qwen-72B-Chat	Alibaba	Foundation		Open-source	72	Unknown
WizardLM	Microsoft	Fine-tuned	LLaMA2-70B	Open-source	70	Aug-23
Tulu-2-DPO	AllenAI/UW	Fine-tuned	LLaMA2-70B	Open-source	70	Nov-23
LLaMA-20-70B-SteerLM-Chat	Nvidia	Fine-tuned	LLaMA2-70B	Open-source	70	Nov-23
Platypus2-70B-instruct	garage-bAInd	Fine-tuned	LLaMA2-70B	Open-source	70	Unknown
PPLX-70B	Perplexity AI	Fine-tuned	Mistral 7B, LLaMA2-70B	Proprietary	70	Real-time updates
llama-2-70b-chat	Meta AI	Foundation		Open-source	70	Jul-23
Deepseek	DeepSeek AI	Foundation		Open-source	67	Nov-23
mixtral-8x7b-instruct-v0.	Mistral AI	Foundation		Open-source	56	Dec-23
Command-Nightly	Cohere	Foundation		Open-source	52	Real-time updates
Claude 2.1	Anthropic	Foundation		Proprietary	52	Unknown
falcon-40b-instruct	TII	Foundation		Open-source	40	Sep-23
Yi-34B-Chat	01.AI	Foundation		Open-source	34	Jun-23
vicuna-33b-v1.3	LargeModel Systems	Fine-tuned	LLaMA	Open-source	33	Aug-23
orca-2-13b	Microsoft	Fine-tuned	LLaMA2-13B	Open-source	13	Oct-23
Baichuan2	Baichuan	Foundation		Open-source	13	Jun-23
Solar	upstage AI	Fine-tuned	LLaMA2	Open-source	11	Nov-23
openchat-3.5-1210	OpenChat	Fine-tuned	Mistral 7B, LLaMA2 7b	Open-source	7	Nov-23
OpenHermes-2p5-Mistral-7B	Teknium	Fine-tuned	Mistral 7B	Open-source	7	Nov-23
zephyr-7b-beta	HuggingFaceH4	Fine-tuned	Mistral 7B	Open-source	7	Oct-23
starling-lm-7b-alpha	UC Berkeley	Fine-tuned	OpenChat 3.5	Open-source	7	Nov-23
StripedHyena-Hessian-7B	hessian.AI, together.AI	Foundation		Open-source	7	Dec-23
alpaca-7b	Stanford CRFM	Fine-tuned	LLaMA-7B	Proprietary	7	Mar-23
RedPajama-INCITE-7B-Chat	together.ai	Fine-tuned	LLaMA2	Open-source	7	Unknown
chatglm3-6b	Zhipu AI, Tsinghua KEG	Foundation		Open-source	6	Oct-23

Note: For proprietary LLMs, size and information cut-offs are not always reported by the developer. In these cases, we triangulate from various estimates.

Table 2: Summary statistics

	N	Mean	Median	SD	Min	Max
Panel A: Implementability						
No. assets	2,048	4.81	5.00	3.64	1.0	73.0
Data available	1,561	0.93	1.00	0.20	0.0	1.0
Some response error	2,048	0.47	0.00	0.50	0.0	1.0
By type						
Weights do not add up to 100%	2,048	0.34	0.00	0.47	0.0	1.0
No recommendation	2,048	0.17	0.00	0.37	0.0	1.0
No portfolio weights	2,048	0.07	0.00	0.25	0.0	1.0
Wrong ticker	2,048	0.05	0.00	0.16	0.0	1.0
No ticker provided	2,048	0.04	0.00	0.18	0.0	1.0
Panel B: Exposure						
Equity	1,564	0.67	0.72	0.27	0.0	1.0
Fixed income	1,564	0.30	0.24	0.25	0.0	1.0
Alternative assets	1,564	0.02	0.00	0.06	0.0	1.0
Cash	1,564	0.01	0.00	0.07	0.0	1.0
Individual stocks	2,048	0.02	0.00	0.10	0.0	1.0
US securities	1,561	0.65	0.70	0.27	0.0	1.0
German securities	1,561	0.04	0.01	0.10	0.0	1.0
Chinese securities	1,561	0.07	0.01	0.19	0.0	1.0
Developed markets	1,561	0.88	0.94	0.20	0.0	1.0
Emerging markets	1,561	0.12	0.06	0.20	0.0	1.0
Domestic securities	1,264	0.49	0.59	0.37	0.0	1.0
Monthly volatility (%)	1,559	2.97	2.98	1.35	0.2	22.1
FF6 market beta	1,559	0.56	0.57	0.25	-0.1	1.3
Idiosyncratic volatility (%)	1,559	0.78	0.86	0.22	-0.1	1.0
Panel C: Performance						
Excess return (%)	1,559	0.36	0.33	0.34	-0.4	3.4
Annual Sharpe ratio	1,559	0.35	0.41	0.33	-3.0	1.2
FF6 alpha (%)	1,559	-0.09	-0.13	0.23	-0.9	3.1
Total expense ratio (%)	1,534	0.18	0.12	0.21	0.0	3.1

Note: Summary statistics are displayed at the LLM–profile level (32 LLMs × 64 profiles). Data available is defined as the share of portfolio assets for which pricing data is available through *YahooFinance*. Some response error is equal to 1 if at least one of the 5 error types occurs in an LLM–profile combination. Alternative assets include commodities, private equity, private debt, cryptocurrencies and real estate (including REITs). Developed markets follow the MSCI definition plus Luxembourg and Liechtenstein. Emerging market follow the MSCI definition. Idiosyncratic volatility is computed as the standard deviation of residuals of a FF6 regression using historical monthly price data from Bloomberg (01/2010–12/2023). Total expense ratios are computed as weighted averages of product-level ratios.

Table 3: Implementability determinants

	(1) No. assets	(2) No. assets	(3) Data available	(4) Data available	(5) Some resp. error	(6) Some resp. error
Model characteristics						
Foundation model	0.467***	0.467***	-0.012	-0.012	-0.073***	-0.073***
	(0.157)	(0.157)	(0.009)	(0.009)	(0.023)	(0.023)
Size > 60B	0.564***	0.564***	-0.010	-0.009	-0.134***	-0.134***
	(0.141)	(0.140)	(0.011)	(0.010)	(0.023)	(0.023)
Open-source	-0.011	-0.011	0.055***	0.059***	0.140***	0.140***
	(0.149)	(0.147)	(0.014)	(0.012)	(0.026)	(0.026)
Investor characteristics						
Age = 45		-0.166		-0.008		-0.007
		(0.186)		(0.013)		(0.028)
Age = 60		-0.176		-0.016*		0.026
		(0.195)		(0.010)		(0.024)
Risk tolerance = high		0.852***		0.020**		-0.008
		(0.159)		(0.009)		(0.021)
Home country = US		0.018		0.009		-0.026
		(0.194)		(0.006)		(0.030)
Home country = Germany		0.120		-0.209***		0.112***
		(0.232)		(0.018)		(0.035)
Home country = China		-0.031		-0.132***		0.089**
		(0.206)		(0.017)		(0.036)
Sustainability = Yes		0.155		0.005		0.100***
		(0.151)		(0.011)		(0.023)
Constant	4.307***	3.906***	0.905***	0.956***	0.477***	0.409***
	(0.207)	(0.294)	(0.016)	(0.016)	(0.031)	(0.044)
Obs.	2,048	2,048	1,561	1,561	2,048	2,048
Model	OLS	OLS	OLS	OLS	OLS	OLS
Adj. R2	0.010	0.021	0.022	0.214	0.060	0.085

Note: Dependent variables are the number of portfolio assets, the share of assets for which historical price data is available on *YahooFinance*, and the incidence of response errors. Omitted categories are "fine-tuned" for LLM type, "≤ 60B" for LLM size, "proprietary" for license type, "30" for age, "low" for risk tolerance, "not specified" for home country, and "no" for sustainability. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$).

Table 4: Summary statistics of LLM-generated vs. robo-advisory recommendations

	(1)			(2)			(1) - (2)	
	LLMs			Robo-advisor				
	N	Mean	SD	N	Mean	SD	Δ	z-score
Exposure								
Equity share	599	0.67	0.27	480	0.59	0.29	0.08	5.14***
Fixed income	599	0.30	0.25	480	0.38	0.27	-0.08	-4.85***
Alternative assets	599	0.02	0.06	480	0.01	0.01	0.01	-14.12***
Cash	599	0.01	0.06	480	0.02	0.04	-0.01	-25.03***
US securities	582	0.65	0.26	480	0.53	0.21	0.12	8.69***
German securities	582	0.06	0.15	480	0.03	0.03	0.03	-10.30***
Developed markets	598	0.92	0.10	480	0.85	0.10	0.07	14.20***
Emerging markets	598	0.08	0.10	480	0.11	0.09	-0.03	–10.29***
Domestic assets	598	0.43	0.37	480	0.38	0.34	0.05	0.25
Domestic equity	582	0.43	0.38	480	0.31	0.30	0.12	2.65**
Monthly volatility (%)	598	2.93	1.35	480	2.03	0.97	0.90	13.12***
FF6 market beta	598	0.56	0.24	476	0.31	0.20	0.26	16.62***
Idiosyncratic volatility (%)	598	0.80	0.19	476	0.55	0.25	0.24	17.54***
Performance								
Excess return (%)	598	0.35	0.31	480	0.19	0.22	0.16	10.18***
Sharpe ratio	598	0.35	0.30	480	0.24	0.29	0.11	7.54***
FF6 alpha (%)	598	-0.12	0.21	476	-0.08	0.13	-0.04	-8.87***
Total expense ratio (%)	594	0.18	0.18	480	0.15	0.11	0.03	2.12*

Note: Column (1) reports summary statistics for the LLM-generated recommendations for profiles 13 through 36. Column (2) reports summary statistics for the robo-advisory recommendations (20 advisors × 12 profiles × 2 experience levels). We elicit portfolios for the lowest and highest experience levels. Column (3) reports the difference in means and the z scores for non-parametric rank tests for differences in means ($^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$). Alternative assets include commodities, cryptocurrencies and real estate (including REITs). Developed markets follow the MSCI definition plus Luxembourg and Liechtenstein. Emerging markets follow the MSCI definition. Idiosyncratic volatility is computed as the standard deviation of residuals of a FF6 regression using historical price data from Bloomberg. Total expense ratios are computed as weighted averages of product-level ratios.

Table 5: Exposure determinants

	Asset classes			Markets			Risk		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Equity share	Fixed income	Ind. stocks	Developed markets	US securities	Domestic securities	Monthly vola. (%)	FF6 β_{mkt}	Idios. vola.
Model characteristics									
Foundation model	-0.018	0.017	0.002	0.008	0.031**	0.033**	-0.079	-0.001	-0.017
	(0.012)	(0.012)	(0.005)	(0.009)	(0.014)	(0.015)	(0.077)	(0.012)	(0.011)
Size > 60B	-0.043***	0.051***	0.003	-0.011	-0.022	0.027*	-0.057	0.022**	0.021**
	(0.011)	(0.011)	(0.004)	(0.009)	(0.014)	(0.016)	(0.058)	(0.010)	(0.010)
Open-source	-0.018	0.035***	-0.001	0.047***	0.037**	-0.067***	-0.142*	-0.005	0.003
	(0.014)	(0.014)	(0.005)	(0.010)	(0.016)	(0.018)	(0.077)	(0.013)	(0.012)
Investor characteristics									
Age = 45	-0.049***	0.051***	0.004	-0.004	0.000	0.013	-0.130**	-0.022*	-0.019
	(0.013)	(0.013)	(0.005)	(0.012)	(0.017)	(0.019)	(0.061)	(0.013)	(0.013)
Age = 60	-0.096***	0.082***	0.017***	0.008	0.025*	0.040**	-0.143*	-0.084***	-0.080***
	(0.013)	(0.012)	(0.005)	(0.009)	(0.014)	(0.016)	(0.075)	(0.012)	(0.012)
Risk tolerance = high	0.300***	-0.286***	0.028***	-0.039***	0.002	-0.004	1.287***	0.283***	0.105***
	(0.011)	(0.010)	(0.004)	(0.008)	(0.013)	(0.014)	(0.059)	(0.010)	(0.010)
Home country = US	-0.006	0.012	0.008*	-0.002	-0.006		0.025	0.001	-0.008
	(0.014)	(0.013)	(0.004)	(0.006)	(0.015)		(0.061)	(0.014)	(0.012)
Home country = Germany	-0.008	0.003	0.019***	-0.028***	-0.158***	-0.604***	-0.064	-0.052***	-0.041***
	(0.017)	(0.016)	(0.006)	(0.008)	(0.020)	(0.015)	(0.100)	(0.016)	(0.014)
Home country = China	0.025	-0.040**	0.030***	-0.283***	-0.227***	-0.418***	0.266**	-0.095***	-0.202***
	(0.018)	(0.017)	(0.007)	(0.022)	(0.023)	(0.023)	(0.115)	(0.018)	(0.020)
Sustainability = Yes	0.089***	-0.081***	-0.002	0.002	0.008	-0.011	0.242***	0.015	-0.006
	(0.012)	(0.011)	(0.005)	(0.010)	(0.014)	(0.016)	(0.070)	(0.012)	(0.011)
Constant	0.577***	0.373***	-0.010*	0.926***	0.677***	0.712***	2.439***	0.466***	0.803***
	(0.022)	(0.021)	(0.006)	(0.015)	(0.026)	(0.026)	(0.111)	(0.021)	(0.019)
Obs.	1,564	1,564	2,048	1,561	1,561	1,264	1,559	1,559	1,559
Adj. R2	0.374	0.377	0.029	0.297	0.125	0.542	0.242	0.357	0.189

Note: Developed markets follow the MSCI definition plus Luxembourg and Liechtenstein. Idiosyncratic volatility is computed as the portfolio-level standard deviation of residuals from FF6 regressions. Omitted categories are "fine-tuned" for LLM type, "$\leq$ 60B" for LLM size, "proprietary" for license type, "30" for age, "low" for risk tolerance, "not specified" for home country, and "no" for sustainability. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$).

Table 6: Dominance statistics for exposure regressions

	(1) Equity share			(2) Monthly volatility (%)			(3) FF6 β_{mkt}			(4) Idiosyncratic volatility		
	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI
Panel A: LLMs												
Investor characteristics												
Age = 45	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.01]	0.00	0.02	[0.00, 0.01]
Age = 60	0.02	0.05	[0.01, 0.04]	0.00	0.00	[0.00, 0.01]	0.02	0.05	[0.00, 0.04]	0.04	0.16	[0.01, 0.07]
Risk tolerance = high	0.26	0.65	[0.21, 0.33]	0.20	0.67	[0.12, 0.30]	0.28	0.70	[0.23, 0.35]	0.06	0.28	[0.03, 0.10]
Sustainability = yes	0.03	0.09	[0.01, 0.06]	0.01	0.02	[0.00, 0.02]	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]
LLM FE	0.08	0.21	[0.05, 0.08]	0.09	0.31	[0.06, 0.10]	0.10	0.25	[0.07, 0.11]	0.12	0.53	[0.05, 0.14]
Obs.		599			598			598			598	
R2		0.40			0.30			0.41			0.22	
Panel B: Robo-advisors												
Investor characteristics												
Age = 45	0.01	0.01	[0.00, 0.01]	0.00	0.00	[0.00, 0.01]	0.00	0.00	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]
Age = 60	0.07	0.09	[0.05, 0.10]	0.03	0.04	[0.02, 0.05]	0.03	0.04	[0.02, 0.05]	0.01	0.01	[0.00, 0.02]
Risk tolerance = high	0.55	0.72	[0.51, 0.61]	0.28	0.34	[0.24, 0.33]	0.25	0.30	[0.22, 0.30]	0.04	0.05	[0.02, 0.06]
Sustainability = yes	0.00	0.00	[0.00, 0.00]	0.01	0.01	[0.00, 0.03]	0.01	0.01	[0.00, 0.02]	0.00	0.00	[0.00, 0.00]
Experience = yes	0.01	0.01	[0.00, 0.03]	0.01	0.01	[0.00, 0.02]	0.00	0.01	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]
Advisor FE	0.13	0.16	[0.08, 0.14]	0.50	0.60	[0.44, 0.54]	0.52	0.64	[0.46, 0.57]	0.80	0.94	[0.75, 0.84]
Obs.		480			480			476			476	
R2		0.76			0.83			0.82			0.85	

Note: The table reports dominance statistics for regressions of equity share (column 1), volatility (column 2), FF6 market beta (column 3), and idiosyncratic volatility (column 4) on the various investor characteristics and LLM/advisor fixed effects. Panel A reports the weights for LLM-generated recommendations for investor profiles 12 through 36 (US and German investors). Panel B reports the weights for robo-advisory recommendations for the same profiles. Confidence intervals are based on 10,000 bootstrap samples. To ensure comparability between LLMs and robo-advisors, we exclude the home country investor characteristic from the analysis, as robo-advisors are exclusively based in either the US or Germany. In unreported results, we find that the main results for LLM portfolios are not affected by this.

Table 7: Sensitivity of exposure to risk tolerance, by LLM type and size

	(1) Equity share	(2) Equity share	(3) Monthly volatility (%)	(4) Monthly volatility (%)	(5) FF6 β_{mkt}	(6) FF6 β_{mkt}	(7) Idiosyncratic volatility	(8) Idiosyncratic volatility
Risk tolerance = high	0.257***	0.261***	1.104***	1.186***	0.244***	0.256***	0.074***	0.097***
	(0.015)	(0.015)	(0.108)	(0.091)	(0.015)	(0.014)	(0.015)	(0.015)
Model type								
Risk tolerance = high × foundation model	0.071***		0.303**		0.065***		0.052***	
	(0.021)		(0.128)		(0.020)		(0.020)	
Model size								
Risk tolerance = high × Size > 60B		0.075***		0.198*		0.053***		0.017
		(0.021)		(0.120)		(0.020)		(0.020)
Constant	0.599***	0.595***	2.531***	2.484***	0.486***	0.478***	0.819***	0.807***
	(0.023)	(0.023)	(0.120)	(0.114)	(0.021)	(0.021)	(0.020)	(0.020)
Obs.	1,564	1,564	1,559	1,559	1,559	1,559	1,559	1,559
Adj. R2	0.378	0.379	0.245	0.243	0.361	0.359	0.192	0.189
Model controls	✓	✓	✓	✓	✓	✓	✓	✓
Profile controls	✓	✓	✓	✓	✓	✓	✓	✓

Note: Model controls include a dummy variable indicating a foundation model (vs. fine-tuned), a dummy variable indicating an open source-LLM (vs. proprietary), and a dummy variable indicating LLM size (> 60B parameters). Profile controls include the investor characteristics listed in Table 5. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$).

Table 8: Performance comparison (LLMs vs. robo-advisors)

$\mathbb{1}(LLM)$	Excess return (%)	Sharpe ratio	FF6 α (%)
Full sample	0.159***	0.107***	-0.046***
	(0.014)	(0.016)	(0.010)
Sub-samples			
Risk tolerance = high	0.172***	0.060***	-0.034**
	(0.022)	(0.020)	(0.016)
Risk tolerance = low	0.145***	0.156***	-0.058***
	(0.015)	(0.025)	(0.012)
Sustainability preference = yes	0.178***	0.145***	-0.038***
	(0.019)	(0.022)	(0.014)
Sustainability preference = no	0.140***	0.070***	-0.056***
	(0.019)	(0.023)	(0.015)
Age = 30	0.137***	0.098***	-0.056***
	(0.024)	(0.026)	(0.016)
Age = 45	0.144***	0.089***	-0.056***
	(0.024)	(0.028)	(0.016)
Age = 60	0.194***	0.134***	-0.032*
	(0.022)	(0.028)	(0.019)
Home country = United States	0.205***	0.175***	-0.017
	(0.019)	(0.023)	(0.014)
Home country = Germany	0.110***	0.036*	-0.075***
	(0.019)	(0.021)	(0.014)
Experience = high (robo-advisors)	0.144***	0.091***	-0.051***
	(0.016)	(0.020)	(0.011)
Experience = low (robo-advisors)	0.173***	0.123***	-0.042***
	(0.016)	(0.020)	(0.012)

Note: The table reports the coefficients for regressions of the performance measures on a dummy variable indicating that a portfolio has been recommended by an LLM (rather than a robo-advisor). We control for all investor characteristics in the full sample and all investor characteristics except the one identifying the respective sub-sample in the sub-samples. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$). The full regression outputs are reported in Table A22 in the Appendix.

Table 9: Determinants of portfolio performance and fees

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Excess	Excess	Sharpe	Sharpe	FF6 α	FF6 α	Total exp.	Total exp.
	return (%)	return (%)	ratio	ratio	(%)	(%)	ratio (%)	ratio (%)
Model characteristics								
Foundation model	0.004	-0.004	-0.020	-0.041*	0.015	0.022	-0.007	0.018
	(0.019)	(0.022)	(0.017)	(0.022)	(0.015)	(0.016)	(0.011)	(0.015)
Size > 60B	-0.004	-0.024	-0.019	-0.047***	-0.014	-0.013	-0.021**	0.008
	(0.014)	(0.017)	(0.015)	(0.016)	(0.010)	(0.012)	(0.010)	(0.010)
Open-source	-0.018	0.018	-0.001	0.031	-0.044***	-0.013	-0.024**	0.009
	(0.020)	(0.025)	(0.019)	(0.022)	(0.016)	(0.020)	(0.011)	(0.012)
Cutoff within 6 months		-0.019		-0.063***		0.016		0.046***
		(0.018)		(0.021)		(0.012)		(0.014)
Investor characteristics								
Age = 45	-0.023	-0.007	-0.026	-0.009	-0.005	0.004	0.009	-0.004
	(0.018)	(0.019)	(0.019)	(0.020)	(0.013)	(0.013)	(0.013)	(0.014)
Age = 60	-0.032*	-0.023	-0.062***	-0.051**	0.029**	0.036**	0.014	0.016
	(0.019)	(0.021)	(0.018)	(0.020)	(0.014)	(0.016)	(0.012)	(0.014)
Risk tolerance = high	0.291***	0.298***	0.210***	0.214***	0.084***	0.078***	0.003	-0.000
	(0.015)	(0.017)	(0.015)	(0.017)	(0.011)	(0.013)	(0.010)	(0.012)
Home country = US	0.026	0.013	0.011	0.003	0.020*	0.012	-0.001	0.003
	(0.017)	(0.019)	(0.018)	(0.020)	(0.012)	(0.013)	(0.010)	(0.011)
Home country = Germany	-0.100***	-0.117***	-0.132***	-0.150***	-0.045***	-0.063***	0.084***	0.068***
	(0.021)	(0.023)	(0.022)	(0.025)	(0.015)	(0.016)	(0.015)	(0.013)
Home country = China	-0.117***	-0.134***	-0.207***	-0.206***	0.047**	0.024	0.141***	0.130***
	(0.030)	(0.034)	(0.026)	(0.026)	(0.023)	(0.027)	(0.018)	(0.021)
Sustainability = Yes	-0.011	-0.026	-0.005	-0.019	0.003	-0.010	0.058***	0.063***
	(0.017)	(0.018)	(0.016)	(0.017)	(0.012)	(0.014)	(0.010)	(0.011)
Constant	0.276***	0.273***	0.357***	0.395***	-0.129***	-0.160***	0.142***	0.057***
	(0.028)	(0.040)	(0.026)	(0.036)	(0.020)	(0.029)	(0.016)	(0.021)
Obs.	1,559	1,186	1,559	1,186	1,559	1,186	1,534	1,166
Adj. R2	0.215	0.238	0.172	0.194	0.058	0.052	0.100	0.092

Note: Omitted categories are "fine-tuned" for LLM type, "$\leq$ 60B" for LLM size, "proprietary" for license type, "30" for age, "low" for risk tolerance, "not specified" for home country, and "no" for sustainability. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1$,$^{**} p < 0.05$,$^{***} p < 0.01$).

Table 10: Domestic allocation and home bias

	All asset classes			Equity only				Home bias			
Home country	Mean	N	SD	d_i	N	SD	m_i	$d_i - m_i$	t ($d_i = m_i$)	$\frac{d_i - m_i}{m_i}$	$\frac{d_i - m_i}{1 - m_i}$
Panel A: All observations											
China	0.29	274	0.36	0.23	271	0.31	0.02	0.21	11.13***	8.69	0.21
Germany	0.11	287	0.20	0.09	277	0.18	0.02	0.07	6.18***	3.40	0.07
United States (all profiles)	0.72	703	0.22	0.73	682	0.24	0.63	0.11	11.56***	0.17	0.28
United States (base profiles)	0.72	311	0.21	0.73	305	0.23	0.63	0.11	8.07***	0.17	0.29
Panel B: Only portfolios without response errors											
China	0.21	178	0.32	0.18	177	0.29	0.02	0.16	7.31***	6.61	0.16
Germany	0.06	186	0.11	0.06	177	0.17	0.02	0.04	3.38***	2.15	0.04
United States (all profiles)	0.71	621	0.22	0.73	606	0.24	0.63	0.10	10.47***	0.16	0.27
United States (base profiles)	0.71	257	0.21	0.73	254	0.23	0.63	0.11	7.26***	0.17	0.28
Panel C: Only high risk-tolerance profiles											
China	0.29	140	0.35	0.25	140	0.32	0.02	0.22	8.31***	9.23	0.23
Germany	0.08	147	0.17	0.06	145	0.11	0.02	0.04	3.74***	1.79	0.04
United States (all profiles)	0.72	356	0.20	0.72	354	0.21	0.63	0.09	8.31***	0.15	0.25
United States (base profiles)	0.73	159	0.20	0.73	158	0.21	0.63	0.10	5.98***	0.16	0.27

Note: The table reports the portfolio weights assigned to domestic securities by home country of the investor profile. For the US, base profiles refer to profiles 13 through 24 (which are directly comparable to the German and Chinese profiles). The first (second) column reports the average domestic allocation d_i, the number of portfolios, and the standard deviation of the domestic allocation for different asset classes (only within the equity portion of a portfolio). The third column reports the respective country weight m_i in the market portfolio as proxied by the MSCI All-Country World Investable Markets Index, as well as three home bias measures (simple weight-gap, weight-gap divided by benchmark weight, weight-gap divided benchmark international weight, cf. Cooper et al., 2018). The table further reports the test statistics of one-sample t-tests for equality of the domestic equity weight d_i and the benchmark weight m_i. Significance is indicated by stars ($^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$). Panel A uses all observations. Panel B only uses portfolio recommendations without erroneous security tickers. Panel C only uses portfolio recommendations for high-risk-tolerance investor profiles.

Figure 1: Theoretical framework

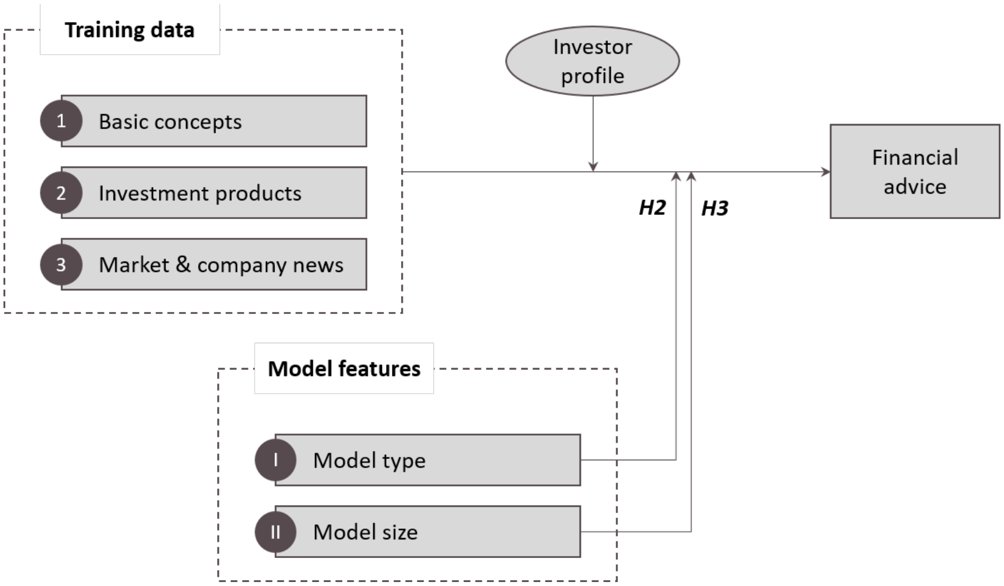
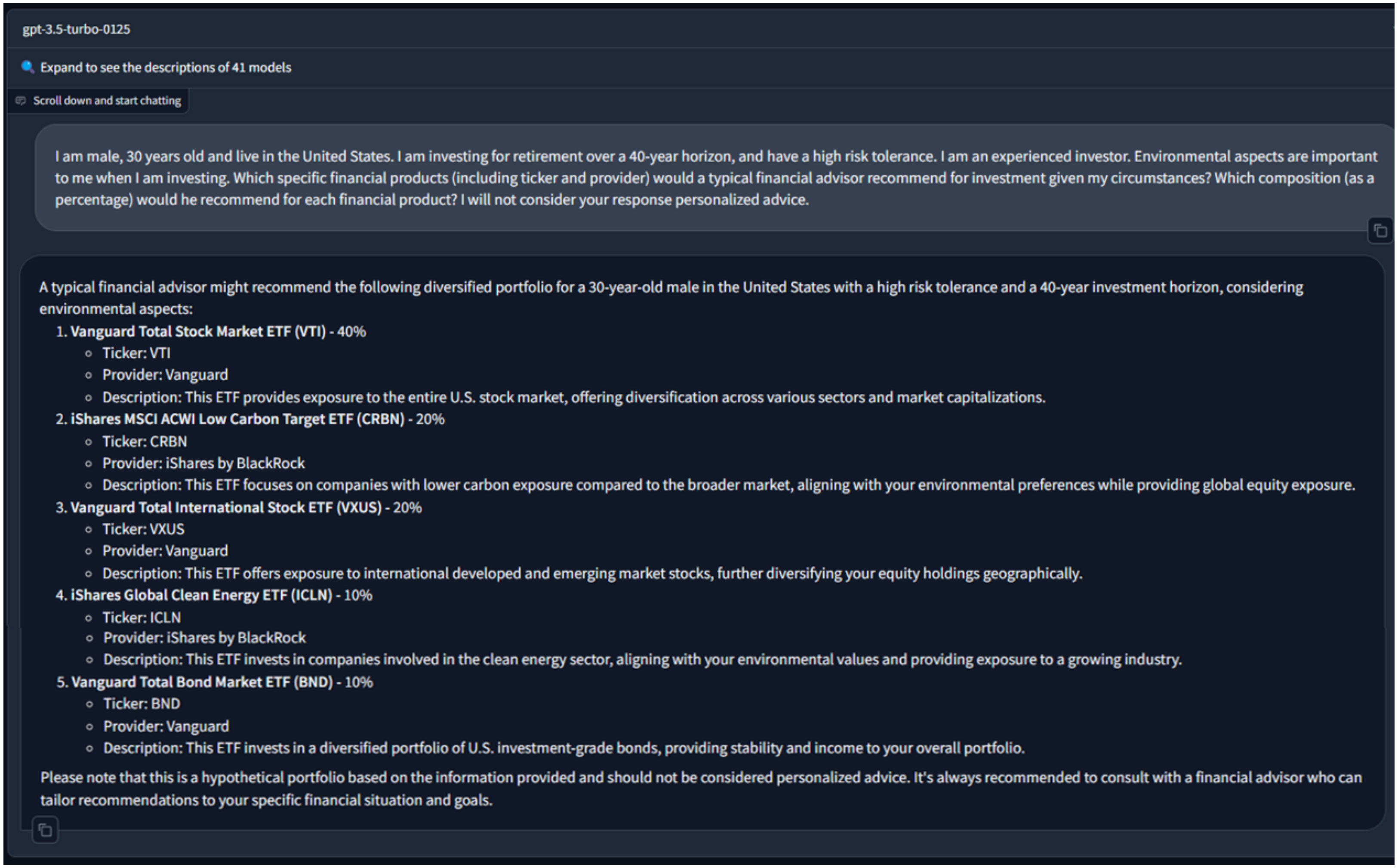


Figure 2: GPT-3.5 Turbo portfolio suggestion (illustration)


gpt-3.5-turbo-0125
Expand to see the descriptions of 41 models
Scroll down and start chatting

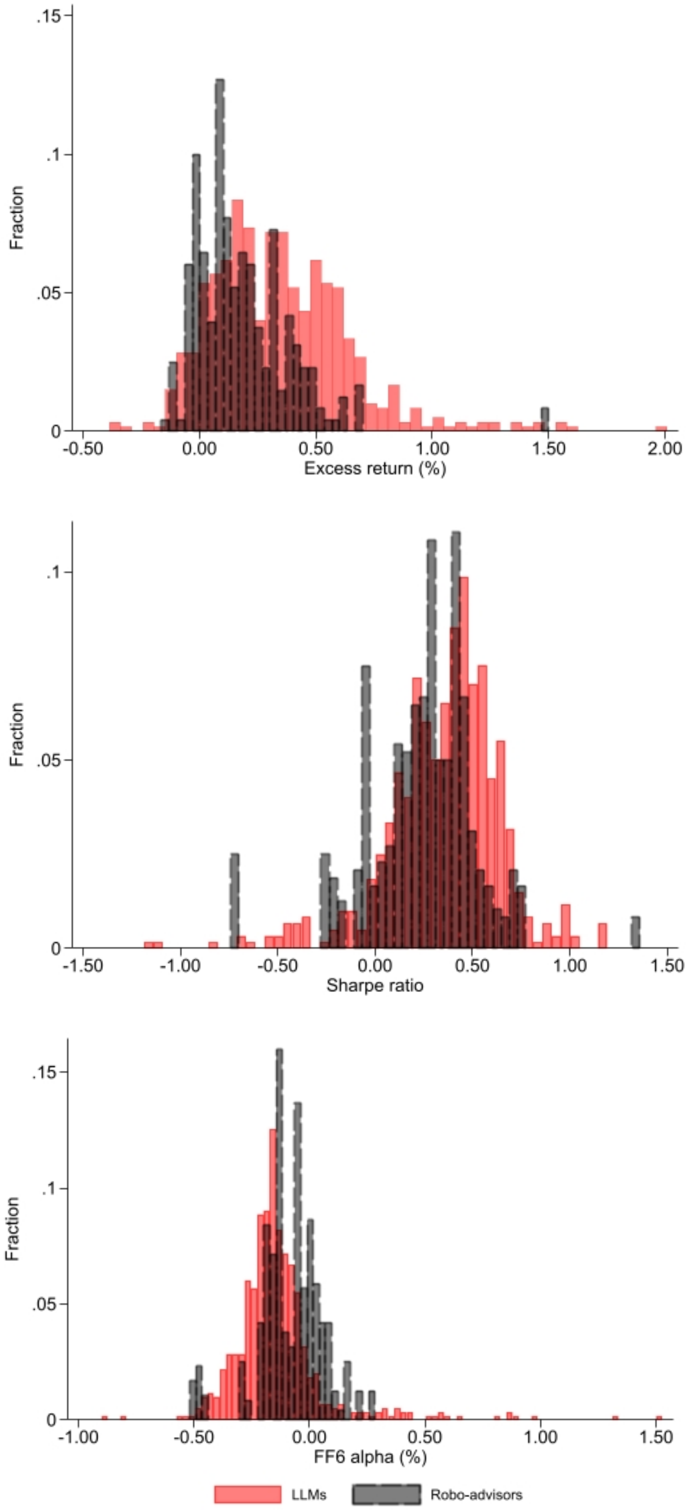
I am male, 30 years old and live in the United States. I am investing for retirement over a 40-year horizon, and have a high risk tolerance. I am an experienced investor. Environmental aspects are important to me when I am investing. Which specific financial products (including ticker and provider) would a typical financial advisor recommend for investment given my circumstances? Which composition (as a percentage) would he recommend for each financial product? I will not consider your response personalized advice.

A typical financial advisor might recommend the following diversified portfolio for a 30-year-old male in the United States with a high risk tolerance and a 40-year investment horizon, considering environmental aspects:

1. **Vanguard Total Stock Market ETF (VTI)** - 40%
   - Ticker: VTI
   - Provider: Vanguard
   - Description: This ETF provides exposure to the entire U.S. stock market, offering diversification across various sectors and market capitalizations.
2. **iShares MSCI ACWI Low Carbon Target ETF (CRBN)** - 20%
   - Ticker: CRBN
   - Provider: iShares by BlackRock
   - Description: This ETF focuses on companies with lower carbon exposure compared to the broader market, aligning with your environmental preferences while providing global equity exposure.
3. **Vanguard Total International Stock ETF (VXUS)** - 20%
   - Ticker: VXUS
   - Provider: Vanguard
   - Description: This ETF offers exposure to international developed and emerging market stocks, further diversifying your equity holdings geographically.
4. **iShares Global Clean Energy ETF (ICLN)** - 10%
   - Ticker: ICLN
   - Provider: iShares by BlackRock
   - Description: This ETF invests in companies involved in the clean energy sector, aligning with your environmental values and providing exposure to a growing industry.
5. **Vanguard Total Bond Market ETF (BND)** - 10%
   - Ticker: BND
   - Provider: Vanguard
   - Description: This ETF invests in a diversified portfolio of U.S. investment-grade bonds, providing stability and income to your overall portfolio.

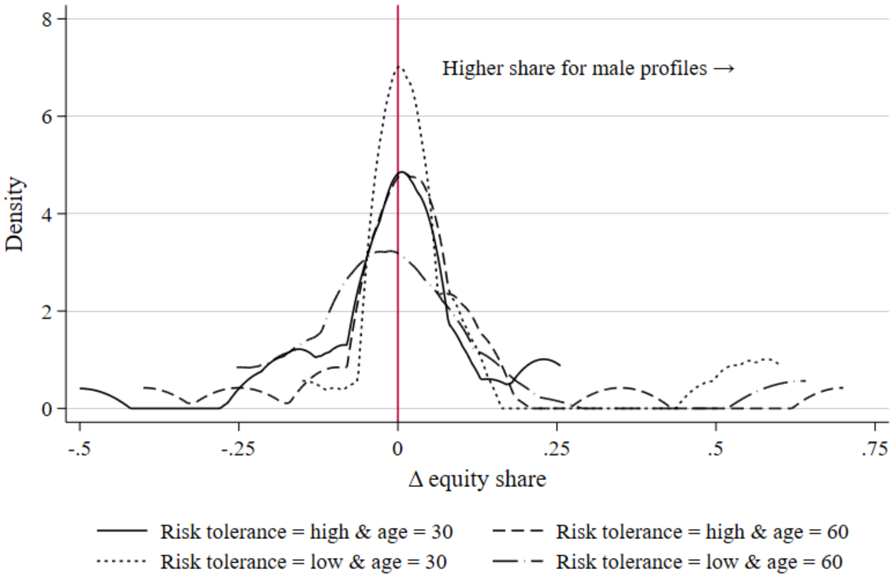
Please note that this is a hypothetical portfolio based on the information provided and should not be considered personalized advice. It's always recommended to consult with a financial advisor who can tailor recommendations to your specific financial situation and goals.

Figure 3: Performance distributions, by advice source



Note: The figures display the distributions of average monthly excess returns (%), annual Sharpe ratios and FF6 αs (%) of portfolio recommendations for profiles 13 to 36, comparing LLMs (solid red bars) and robo-advisors (dashed black bars). For expositional purposes, we use unadjusted FF6 α point estimates (i.e., not accounting for statistical significance).

Figure 4: Gender difference in equity share



Note: The figures display density plots of the difference in equity shares in portfolios recommended to male versus female investors. Positive values indicate that the equity share recommended to a male profile was higher than the equity share recommended to the equivalent (in terms of risk tolerance and age) female profile.

Online Appendix:
Using large language models for financial advice

1. Tables

Table A1: Availability of information in CommonCrawl dataset

Information type	**Corresponding domains in CommonCrawl**
Basic concepts	Educational institutions (yale.edu, duke.edu, berkeley.edu) General-purpose knowledge domains (en-academic.com, wikipedia.org) Finance-specific domains (investing.com)
Investment products	Financial analysis domains (tradingview.com, yahoo.com)
Market & company news	News domains (euronews.com, wikinews.com)

Table A2: Investor profiles

#	Age	Investment horizon	Risk tolerance	Home country	Sustain. preference	Gender	Experience
1	30	40	High	-	No	-	-
2	30	40	High	-	Yes	-	-
3	30	40	Low	-	No	-	-
4	30	40	Low	-	Yes	-	-
5	45	15	High	-	No	-	-
6	45	15	High	-	Yes	-	-
7	45	15	Low	-	No	-	-
8	45	15	Low	-	Yes	-	-
9	60	5	High	-	No	-	-
10	60	5	High	-	Yes	-	-
11	60	5	Low	-	No	-	-
12	60	5	Low	-	Yes	-	-
13	30	40	High	United States	No	-	-
14	30	40	High	United States	Yes	-	-
15	30	40	Low	United States	No	-	-
16	30	40	Low	United States	Yes	-	-
17	45	15	High	United States	No	-	-
18	45	15	High	United States	Yes	-	-
19	45	15	Low	United States	No	-	-
20	45	15	Low	United States	Yes	-	-
21	60	5	High	United States	No	-	-
22	60	5	High	United States	Yes	-	-
23	60	5	Low	United States	No	-	-
24	60	5	Low	United States	Yes	-	-
25	30	40	High	Germany	No	-	-
26	30	40	High	Germany	Yes	-	-
27	30	40	Low	Germany	No	-	-
28	30	40	Low	Germany	Yes	-	-
29	45	15	High	Germany	No	-	-
30	45	15	High	Germany	Yes	-	-
31	45	15	Low	Germany	No	-	-
32	45	15	Low	Germany	Yes	-	-
33	60	5	High	Germany	No	-	-
34	60	5	High	Germany	Yes	-	-
35	60	5	Low	Germany	No	-	-
36	60	5	Low	Germany	Yes	-	-
37	30	40	High	China	No	-	-
38	30	40	High	China	Yes	-	-
39	30	40	Low	China	No	-	-
40	30	40	Low	China	Yes	-	-
41	45	15	High	China	No	-	-
42	45	15	High	China	Yes	-	-
43	45	15	Low	China	No	-	-

44	45	15	Low	China	Yes	-	-
45	60	5	High	China	No	-	-
46	60	5	High	China	Yes	-	-
47	60	5	Low	China	No	-	-
48	60	5	Low	China	Yes	-	-
49	30	40	High	United States	No	Male	-
50	30	40	Low	United States	No	Male	-
51	60	5	High	United States	No	Male	-
52	60	5	Low	United States	No	Male	-
53	30	40	High	United States	No	Female	-
54	30	40	Low	United States	No	Female	-
55	60	5	High	United States	No	Female	-
56	60	5	Low	United States	No	Female	-
57	30	40	High	United States	No	-	Experienced
58	30	40	Low	United States	No	-	Experienced
59	60	5	High	United States	No	-	Experienced
60	60	5	Low	United States	No	-	Experienced
61	30	40	High	United States	No	-	Not experienced
62	30	40	Low	United States	No	-	Not experienced
63	60	5	High	United States	No	-	Not experienced
64	60	5	Low	United States	No	-	Not experienced

Table A3: Pairwise correlations of LLM characteristics

	(1)	(2)	(3)	(4)
(1) Foundation model	1.000			
(2) Open-source	-0.323*	1.000		
(3) Size > 60B	0.143	-0.357**	1.000	
(4) Cutoff within 6 months	-0.220	0.164	-0.132	1.000

Note: The table reports pairwise correlations of LLM characteristics (N = 32). Stars indicate significance levels ($^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$).

Table A4: Robo-advisor overview

Robo-advisor	**AuM (in million)**	**Service fees p.a.**	**Minimum investment**	**Risk profiles**	**Asset classes**
Panel A: US Robo-Advisors					
Vanguard Digital Advisor	$ 99,711	0.10% - 0.30%	$100	101	Equity, fixed income, money market
Wealthfront	$ 77,070	0.25%	$500	20	Equity, fixed income
Schwab Intelligent Portfolios	$ 65,885	$0	$500	12	Equity, fixed income, cash
E*Trade	$ 18,200	0.30%	$500	6	Equity, fixed income
Acorns	$ 8,211	$36 - $144	$5	5	Equity, fixed income
SigFig	$ 2,797	0% - 0.25%	$2,000	20	Equity, fixed income
Fidelity Go	$ 1,500*	0% - 0.35%	$10	8	Equity, fixed income, money market
SoFi	$ 1,365	0.25%	$50	5	Equity, fixed income, cash
Ally Invest	$ 1,128	0.30%	$100	5	Equity, fixed income, cash
Empower	$ 412	0.20% - 0.50%	$5,000	5	Equity, fixed income, cash
Panel B: German Robo-Advisors					
Scalable Wealth	€ 4,000*	0.75%	€20	8 - 11	Equity, fixed income, commodities, money market
Quirion	€ 1,500*	0.48% - 1.20%	€25	11	Equity, fixed income, cash
Weltsparen by Raisin	€ 1,000*	0.46%	€500	5	Equity, fixed income
Growney	€ 500*	0.68%	€500	5	Equity, fixed income
Ginmon	€ 300*	0.75%	€1,000	10	Equity, fixed income, commodities, real estate
bevestor	€ 299	0.80% - 1%	€500	5	Equity, fixed income, commodities, money market
Whitebox	€ 150*	0.35%	€50	3	Equity, fixed income, cash
OSKAR	€ 100*	0.80% - 1%	€25	6	Equity, fixed income, commodities
VisualVest	€ 100*	0.60%	€500	7	Equity, fixed income, commodities, money market
fintego	€ 50*	0.70%	€2,500	5	Equity, fixed income, money market

Note: Assets under Management were obtained from Form ADV filings for US robo-advisors. For E*Trade (Morgan Stanley), Schwab Intelligent Portfolios (Schwab), and Vanguard Digital Advisor (Vanguard), we report the discretionary AuM figures. AuM values marked with an asterisk (*) represent estimates derived from industry reports as the actual figures are not disclosed by the robo-advisors.

Table A5: Summary statistics (robo-advisor portfolios)

	(1)			(2)			(1) - (2)	
	United States			Germany				
	N	Mean	SD	N	Mean	SD	Δ	z score
Exposure								
No. assets	240	7,88	3,86	240	9,89	3,40	-2,01	-7.58***
Equity share	240	0,61	0,28	240	0,58	0,29	0,03	0.82
Fixed income	240	0,35	0,26	240	0,40	0,29	-0,05	-1.27
Alternatives	240	0,01	0,01	240	0,01	0,02	0,00	1.37
Cash	240	0,03	0,04	240	0,01	0,03	0,02	7,98***
US securities	240	0,70	0,12	240	0,36	0,12	0,35	18,26***
German securities	240	0,02	0,01	240	0,05	0,03	-0,04	-17.90***
Developed markets	240	0,89	0,08	240	0,81	0,09	0,08	9,17***
Emerging markets	240	0,08	0,07	240	0,14	0,09	-0,06	-8.60***
Domestic securities	240	0,70	0,12	240	0,05	0,03	0,65	18.96***
Domestic equity	240	0,59	0,16	240	0,03	0,01	0,56	17.06***
Monthly volatility (%)	240	2,07	1,11	240	1,99	0,80	0,08	-0.08
FF6 market beta	236	0,29	0,23	240	0,32	0,15	-0,02	-2.89***
Idiosyncratic volatility (%)	236	0,46	0,31	240	0,65	0,09	-0,20	-6.92***
Performance								
Excess return (%)	240	0,19	0,25	240	0,18	0,17	0,02	0.10
Annual Sharpe ratio	240	0,24	0,35	240	0,24	0,22	0,00	0.84
FF6 alpha (%)	236	-0,08	0,15	240	-0,07	0,09	0,00	2.60***
Total expense ratio (%)	240	0,13	0,14	240	0,18	0,05	-0,06	-12.86***

	(1)			(2)			(1) - (2)	
	Experience = low			Experience = high				
	N	Mean	SD	N	Mean	SD	Δ	z-score
Exposure								
No. assets	240	9,14	3,76	240	8,64	3,77	0,50	1,55
Equity share	240	0,56	0,28	240	0,62	0,29	-0,06	-2,47**
Fixed income	240	0,41	0,27	240	0,35	0,28	0,06	2,39**
Alternatives	240	0,01	0,01	240	0,01	0,01	0,00	-0,49
Cash	240	0,02	0,04	240	0,02	0,04	0,00	0,11
US securities	240	0,53	0,21	240	0,53	0,20	0,00	-0,22
German securities	240	0,04	0,03	240	0,03	0,03	0,01	0,28
Developed markets	240	0,85	0,10	240	0,85	0,10	0,00	0,34
Emerging markets	240	0,11	0,09	240	0,12	0,08	-0,01	-1,16
Domestic securities	240	0,38	0,34	240	0,38	0,34	0,00	0.70
Domestic equity	240	0,31	0,30	240	0,31	0,30	0,00	-0,05
Monthly volatility (%)	240	1,96	0,92	240	2,10	1,01	-0,14	-1,42
FF6 market beta	238	0,29	0,19	238	0,32	0,21	-0,03	-1,22
Idiosyncratic volatility (%)	238	0,55	0,25	238	0,56	0,25	-0,01	-0,29
Performance								
Excess return (%)	240	0.17	0.21	240	0.20	0.22	-0.03	-1.72*
Annual Sharpe ratio	240	0.22	0.29	240	0.26	0.29	-0.04	-1.56
FF6 alpha (%)	238	-0.08	0.13	238	-0.07	0.12	-0.01	-1.18
Total expense ratio (%)	240	0.15	0.11	240	0.16	0.11	-0.01	-0.40

Note: Summary statistics are displayed at the advisor-profile level (20 LLMs × 12 profiles × 2 experience levels). We elicit portfolios for the lowest and highest experience levels. Developed markets follow the MSCI definition plus Luxembourg and Liechtenstein. Emerging markets follow the MSCI definition. Idiosyncratic volatility is computed as the standard deviation of residuals of a FF6 regression using historical monthly price data from Bloomberg (01/2010 through 12/2023). Total expense ratios are computed as weighted averages of product-level ratios. Column 3 reports the difference in means and the z score from a non-parametric rank test for differences in means ($^{*}p < 0.1$,** $p < 0.05$,*** $p < 0.01$).

Table A6: Robo-advisory questionnaire

Category	Survey question	Answer
Experience	What is your understanding of stocks, bonds, and ETFs?	highest and lowest
	In which area do you have knowledge?	highest and lowest
	Follow-up question: Experience in years & number of transactions per year	all and none
	If no: knowledge and experience indicated are not yet sufficient. Confirm that you read the following documents	Yes
	How well do you know about investments?	highest and lowest
Investment Amount	Initial contribution	$10,000
	Contribution Amount per Month	$0
	Contribution Frequency	none
Profile	How do you file your taxes?	single
	Estimated retirement age	according to profile
	Time Horizon	according to profile
	Date of Birth	according to profile
	Which account type are you interested in?	taxable
	For which target are you investing	retirement
	Who are you investing for?	myself
	Net disposable income (monthly)	$1,500
	Yearly income before taxes USA	$60,580
	Net worth USA (household)	$193,000
	Disposable assets Germany	€58,000
	Non-disposable assets Germany	€163,000
	Liabilities Germany	€31,565
Risk	When you hear "risk" related to your finances, what is the first thought that comes to mind?	highest and lowest
	Have you ever experienced a 20% or more decline in the value of your investments in one year?	highest and lowest
	What did you do when you experienced a 20% decline in value of your investments?	highest and lowest
	How would you describe your approach to making important financial decisions?	highest and lowest
	How much investment value fluctuation would you be comfortable with 1 year from now?	highest and lowest
	How do you assess your risk tolerance?	highest and lowest
	Which scenario would you be most comfortable with?	highest and lowest
	How would you like to determine your risk tolerance?	highest and lowest
	Is there a chance you might need the money invested in this account to cover large, unexpected expenses?	highest and lowest
	Based on my financial situation, I can weather market downturns and absorb losses without jeopardizing my goal for this account.	highest and lowest
	What's your comfort with risk?	highest and lowest
	When investing my money, I prioritize returns.	highest and lowest
	The risk of losing part of my assets weighs heavily on me.	highest and lowest
	The security of an investment is the most important thing to me.	highest and lowest
	I am reluctant to take risks in financial matters.	highest and lowest
	Even small losses make me nervous.	highest and lowest
	What is fundamentally important to you when investing money?	highest and lowest
	Capital markets are susceptible to fluctuations. What loss in value makes you nervous?	highest and lowest
	How high is your willingness to take risks when investing?	highest and lowest
	Can you bear this predicted loss of capital?	yes and no

Continued on next page

Table A6 – Continued from previous page

Category	Survey question	Answer
	The value of investments can go up or down every year, acceptable for me are:	highest and lowest
	How risky do you want your investment to be?	highest and lowest
	Does this investment make you nervous about major price fluctuations with possible interim losses?	highest and lowest
	Do you want to generate a high profit with this investment and therefore take higher risks?	highest and lowest
	You are investing a portfolio with €1,000.	highest and lowest
	For a higher potential return, you have to accept greater risks of loss. What would you do if you lost 20%?	highest and lowest
Sustainability	Would you like an investment strategy with sustainability features?	according to profile
	Any specific sustainability topics?	nothing specific
Other	Would you like to model options for your retirement goal?	no
	Tax-Loss Harvesting	no
	Municipal bonds	no
	Employment status	full time
	Select the portfolio preference that fits you	globally
	What are you investing for?	no ERISA plan
	Is a retirement account right for you?	no ERISA plan
	Choose the IRA that works for you	no ERISA plan
	When I need this money, I'll withdraw it over a period of	highest and lowest
	How many years from now will you need to start withdrawing funds from this account?	highest and lowest
	In which state do you file your taxes?	California
	I'll estimate my retirement expenses based on: (optional)	no
	First things first: select which Robo Portfolio to start with	market focused
	How long can you live off your 'iron reserves'?	highest and lowest
	Would you like to use our investment protection service?	no
	Do you want to achieve a specific savings goal?	no

Note: The table presents the survey questions and corresponding responses from the robo-advisory questionnaires. The questions are organized into six categories: experience, investment amount, profile, risk, sustainability, and special considerations. Variations in responses for risk tolerance and experience are explicitly displayed in the answer column. For the initial contribution and net disposable income, we use figures from Oehler and Horn (2024). The yearly income before taxes and net worth (household) are from the US Bureau of Labor Statistics (2024) for the United States, and the disposable assets, non-disposable assets, and liabilities are obtained from the Statistisches Bundesamt (Destatis) (2024) for Germany.

Table A7: Pairwise correlations of portfolio characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1) No. assets	1														
(2) Data available	0.0314	1													
(3) Some response error	-0.0141	-0.195***	1												
(4) Equity share	0.0735**	-0.0141	-0.0132	1											
(5) Fixed income	-0.138***	0.0437	-0.0414	-0.938***	1										
(6) Developed markets	-0.0539*	0.346***	-0.130***	-0.135***	0.197***	1									
(7) US securities	0.00825	0.413***	-0.0914***	-0.0670**	0.0804**	0.599***	1								
(8) Domestic securities	-0.0162	0.0932***	-0.0478	-0.0169	0.000792	-0.114***	0.201***	1							
(9) Monthly volatility	0.102***	0.0497*	-0.0621*	0.615***	-0.618***	-0.145***	-0.0433	0.0121	1						
(10) FF6 market beta	0.0638*	0.187***	-0.171***	0.737***	-0.693***	0.134***	0.113***	-0.0317	0.688***	1					
(11) Idiosyncratic volatility	0.00242	0.259***	-0.190***	0.415***	-0.330***	0.486***	0.255***	-0.133***	0.131***	0.677***	1				
(12) Excess return	0.216***	0.291***	-0.0980***	0.471***	-0.476***	0.248***	0.377***	0.0516	0.664***	0.661***	0.380***	1			
(13) Sharpe ratio	0.187***	0.364***	-0.112***	0.456***	-0.405***	0.362***	0.415***	0.0413	0.452***	0.635***	0.613***	0.791***	1		
(14) FF6 alpha	0.247***	0.0962***	0.0227	0.159***	-0.203***	-0.0271	0.259***	0.142***	0.458***	0.126***	-0.159***	0.733***	0.448***	1	
(15) Total expense ratio	0.124***	-0.289***	0.217***	0.0581*	-0.148***	-0.438***	-0.242***	0.00404	0.0120	-0.197***	-0.395***	-0.252***	-0.446***	-0.0206	1

Note: The table reports pairwise correlations of portfolio characteristics (N = 2,048). Stars indicate significance levels ($^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$).

Table A8: Effect of error correction prompts (OLS)

	(1) No PF weights specified	(2) PF weights don't add up to 100%	(3) No specific product	(4) No ticker provided	(5) Wrong ticker
Correction	-0.136***	-0.023	-0.165***	-0.210***	-0.103***
	(0.016)	(0.014)	(0.018)	(0.018)	(0.020)
Obs.	2,128	2,128	2,128	2,128	2,128
Adj. R2	0.052	0.028	0.172	0.155	0.121
Model controls	✓	✓	✓	✓	✓
Profile controls	✓	✓	✓	✓	✓

Note: The table reports the coefficients of OLS regressions of error incidences on a dummy variable indicating whether a correction prompt for the respective error has been included (Correction), as well as LLM and profile characteristics. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

Table A9: Effect of error correction prompts (logistical regression)

	(1) No PF weights specified	(2) PF weights don't add up to 100%	(3) No specific product	(4) No ticker provided	(5) Wrong ticker
Correction	-1.012***	-0.203	-0.980***	-1.211***	-0.475***
	(0.124)	(0.130)	(0.105)	(0.108)	(0.094)
Obs.	2,128	2,128	2,128	2,128	2,128
Pseudo R2	0.063	0.043	0.158	0.144	0.096
Model controls	✓	✓	✓	✓	✓
Profile controls	✓	✓	✓	✓	✓

Note: The table reports the coefficients of logistical regressions of error incidences on a dummy variable indicating whether a correction prompt for the respective error has been included (Correction), as well as LLM and profile characteristics. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

Table A10: Implementability determinants (robustness: logistical regressions & sub-sample)

	Full sample		Without erroneous models				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Data	Some	No.	Data	Some	Data	Some
	available	resp. error	assets	available	resp. error	available	resp. error
Model characteristics							
Foundation model	-0.629	-0.315***	0.510***	-0.018*	-0.109***	-0.717	-0.471***
	(0.511)	(0.098)	(0.166)	(0.009)	(0.025)	(0.514)	(0.104)
Size > 60B	0.306	-0.578***	0.117	-0.015	-0.116***	0.220	-0.508***
	(0.388)	(0.099)	(0.152)	(0.011)	(0.024)	(0.395)	(0.105)
Open-source	1.739***	0.613***	0.163	0.051***	0.078***	1.594***	0.351***
	(0.440)	(0.112)	(0.152)	(0.013)	(0.027)	(0.454)	(0.118)
Investor characteristics							
Age = 45	-0.476	-0.032	-0.183	-0.009	-0.006	-0.478	-0.025
	(0.465)	(0.122)	(0.195)	(0.013)	(0.029)	(0.465)	(0.129)
Age = 60	-0.744*	0.115	-0.154	-0.015	0.020	-0.747*	0.090
	(0.451)	(0.107)	(0.208)	(0.010)	(0.025)	(0.451)	(0.113)
Risk tolerance = high	0.696*	-0.034	0.918***	0.020**	-0.017	0.699*	-0.076
	(0.362)	(0.093)	(0.169)	(0.009)	(0.022)	(0.363)	(0.098)
Home country = US		-0.114	-0.001	0.010*	-0.031		-0.142
		(0.132)	(0.201)	(0.006)	(0.032)		(0.140)
Home country = Germany	-17.100	0.487***	0.112	-0.217***	0.121***	-17.132	0.518***
	(2668.718)	(0.151)	(0.242)	(0.019)	(0.037)	(2724.801)	(0.158)
Home country = China	-16.894	0.384**	-0.083	-0.136***	0.098***	-16.929	0.420***
	(2668.718)	(0.151)	(0.212)	(0.018)	(0.038)	(2724.801)	(0.158)
Sustainability = Yes	0.755**	0.437***	0.158	0.007	0.105***	0.754**	0.454***
	(0.367)	(0.100)	(0.158)	(0.011)	(0.024)	(0.367)	(0.105)
Constant	18.851	-0.406**	4.165***	0.966***	0.429***	19.032	-0.314
	(2668.718)	(0.188)	(0.317)	(0.017)	(0.046)	(2724.801)	(0.198)
Obs.	858	2,048	1,856	1,500	1,856	824	1,856
Adj. R2			0.016	0.221	0.075		
Pseude R2	0.237	0.067				0.235	0.061
Model	Logit	Logit	OLS	OLS	OLS	Logit	Logit

Note: Columns 1 and 2 are estimated using the full sample. Columns 3 through 7 use the full sample excluding LLMs with either 100% response errors or 0% data availability. Dependent variables are the number of portfolio assets, the share of assets for which historical price data is available on YahooFinance, and the incidence of response errors. Omitted categories are "fine-tuned" for LLM type, "≤ 60B" for LLM size, "proprietary" for license type, "30" for age, "low" for risk tolerance, "not specified" for home country, and "no" for sustainability. Standard errors are reported in parentheses ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$) and are heteroskedasticity-robust for the OLS specifications.

Table A11: Implementability determinants (robustness: continuous size specification)

	(1) No. assets	(2) No. assets	(3) Data available	(4) Data available	(5) Some resp. error	(6) Some resp. error
Model characteristics						
Foundation model	0.140	0.140	0.000	0.001	-0.010	-0.010
	(0.148)	(0.147)	(0.011)	(0.010)	(0.024)	(0.024)
ln (size)	0.371***	0.371***	-0.015**	-0.015***	-0.072***	-0.072***
	(0.062)	(0.062)	(0.006)	(0.005)	(0.010)	(0.010)
Open-source	0.350**	0.350**	0.034**	0.036**	0.079***	0.079***
	(0.173)	(0.172)	(0.016)	(0.014)	(0.029)	(0.029)
Investor characteristics						
Age = 45		-0.166		-0.008		-0.007
		(0.186)		(0.013)		(0.028)
Age = 60		-0.176		-0.016		0.026
		(0.195)		(0.010)		(0.024)
Risk tolerance = high		0.852***		0.020**		-0.008
		(0.159)		(0.009)		(0.021)
Home country = US		0.018		0.009		-0.026
		(0.193)		(0.006)		(0.030)
Home country = Germany		0.120		-0.210***		0.112***
		(0.231)		(0.018)		(0.035)
Home country = China		-0.031		-0.132***		0.089**
		(0.204)		(0.017)		(0.036)
Sustainability = Yes		0.155		0.005		0.100***
		(0.150)		(0.011)		(0.023)
Constant	3.104***	2.703***	0.963***	1.017***	0.693***	0.625***
	(0.353)	(0.415)	(0.029)	(0.028)	(0.051)	(0.060)
Obs.	2,048	2,048	1,561	1,561	2,048	2,048
Model	OLS	OLS	OLS	OLS	OLS	OLS
Adj. R2	0.016	0.028	0.026	0.218	0.067	0.092

Note: Dependent variables are the number of portfolio assets, the share of assets for which historical price data is available on YahooFinance, and the incidence of response errors. Omitted categories are "fine-tuned" for model type, "proprietary" for license type, "30" for age, "low" for risk tolerance, "not specified" for home country, and "no" for sustainability. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$).

Table A12: Number of securities recognized by ChatGPT-4o by domicile country and instrument type

	United States	**Germany**	**China**	**Euro area**	**China + Hong Kong**
Equity ETFs	2,600	175	550	1,700	870
Sustainable equity ETFs	150	55	5	425	13
Fixed-income ETFs	1,050	45	17	1,250	55
Stocks	4,250	429	4,350	4,250	4,900

Note: The table reports the number of securities ChatGPT-4o reports to exist by instrument type and domicile country. The prompts were: “How many [instrument type] domiciled in [country] are there?”

Table A13: Summary statistics for alternative factor model specifications

	N	Mean	Median	SD	Min	Max
Panel A: Developed market factors, monthly rebalancing						
Excess return (%)	1,559	0.36	0.33	0.34	-0.4	3.4
Monthly volatility (%)	1,559	2.97	2.98	1.35	0.2	22.1
Annual Sharpe ratio	1,559	0.35	0.41	0.33	-3.0	1.2
FF6 alpha (%)	1,559	-0.09	-0.13	0.23	-0.9	3.1
FF6 market beta	1,559	0.56	0.57	0.25	-0.1	1.3
Idiosyncratic volatility (%)	1,559	0.78	0.86	0.22	-0.1	1.0
Panel B: Region-specific factors, monthly rebalancing						
Excess return (%)	1,559	0.36	0.33	0.34	-0	3
Monthly volatility (%)	1,559	2.97	2.98	1.35	0	22
Annual Sharpe ratio	1,559	0.35	0.41	0.33	-3	1
FF6 alpha (%)	1,559	-0.06	-0.09	0.26	-1	4
FF6 market beta	1,559	0.56	0.57	0.25	-0	1
Idiosyncratic volatility (%)	1,559	0.77	0.83	0.20	-0	1
Panel C: Region-specific factors, annual rebalancing						
Excess return (%)	1,559	0.32	0.27	0.36	-2	5
Monthly volatility (%)	1,559	3.00	2.88	1.87	0	52
Annual Sharpe ratio	1,559	0.33	0.35	0.30	-3	1
FF6 alpha (%)	1,559	-0.10	-0.13	0.29	-1	6
FF6 market beta	1,559	0.55	0.55	0.22	-0	1
Idiosyncratic volatility (%)	1,559	0.76	0.82	0.20	-0	1

Note: Summary statistics are displayed at the LLM–profile level (32 LLMs × 64 profiles). Panel A is the main specification (see Table 2) and uses developed market factor portfolios (provided by Kenneth French) for all portfolios while assuming monthly portfolio rebalancing. Panels B and C use developed market factor portfolios for the German and US profiles and emerging market factor portfolios (both provided by Kenneth French) for Chinese profiles. Panel B assumes monthly portfolio rebalancing; panel C assumes annual portfolio rebalancing. Idiosyncratic volatility is computed as the standard deviation of residuals of a FF6 regression using historical monthly price data from Bloomberg (01/2010 through 12/2023).

Table A14: Exposure determinants (robo-advisory recommendations)

	Asset classes		**Markets**			**Risk**		
	(1)	(2)	(4)	(5)	(6)	(7)	(8)	(9)
	Equity share	Fixed income	Developed markets	US securities	Domestic securities	Monthly volatility (%)	FF6 β_{mkt}	Idiosyncratic volatility
Age = 45	-0.026*	0.026*	0.003	0.004	0.005	-0.053	-0.010	-0.004
	(0.015)	(0.014)	(0.005)	(0.008)	(0.004)	(0.044)	(0.009)	(0.011)
Age = 60	-0.179***	0.175***	0.020***	0.027***	0.039***	-0.406***	-0.083***	-0.050***
	(0.017)	(0.016)	(0.005)	(0.008)	(0.004)	(0.047)	(0.010)	(0.011)
Risk tolerance = high	0.422***	-0.401***	-0.037***	-0.013**	-0.063***	1.021***	0.197***	0.100***
	(0.013)	(0.012)	(0.004)	(0.006)	(0.004)	(0.037)	(0.008)	(0.009)
Sustainability = Yes	-0.007	0.007	0.004	-0.007	-0.006*	-0.200***	-0.031***	-0.003
	(0.013)	(0.012)	(0.004)	(0.006)	(0.004)	(0.037)	(0.008)	(0.009)
Experienced Investor	0.059***	-0.059***	-0.002	0.002	-0.007**	0.141***	0.026***	0.007
	(0.013)	(0.012)	(0.004)	(0.006)	(0.004)	(0.037)	(0.008)	(0.009)
Constant	0.263***	0.647***	0.966***	0.815***	0.840***	1.190***	0.145***	0.283***
	(0.051)	(0.039)	(0.005)	(0.018)	(0.009)	(0.130)	(0.033)	(0.055)
Obs.	480	480	480	480	480	480	476	476
Adj. R2	0.749	0.749	0.762	0.697	0.928	0.820	0.811	0.842
Advisor FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: Developed markets follow the MSCI definition plus Luxembourg and Liechtenstein. Idiosyncratic volatility is computed as the portfolio-level standard deviation of residuals from FF6 regressions. We estimate all regression models using OLS. Omitted categories are "30" for age, "low" for risk tolerance, "no" for sustainability, and "low" for investor experience. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$).

Table A15: Exposure determinants (robustness: implementability)

	Asset classes		**Markets**			**Risk**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Equity share	Fixed income	Developed markets	US securities	Domestic securities	Monthly volatility (%)	FF6 β_mkt	Idiosyncratic volatility
Model characteristics								
Foundation model	-0.020	0.018	0.011	0.037***	0.028**	-0.082	-0.001	-0.017
	(0.012)	(0.012)	(0.008)	(0.013)	(0.014)	(0.076)	(0.011)	(0.011)
Size > 60B	-0.043***	0.053***	-0.009	-0.018	0.026*	-0.061	0.021**	0.021**
	(0.011)	(0.011)	(0.009)	(0.013)	(0.015)	(0.057)	(0.010)	(0.010)
Open-source	-0.011	0.029**	0.034***	0.009	-0.053***	-0.125	-0.002	-0.000
	(0.014)	(0.014)	(0.010)	(0.016)	(0.017)	(0.080)	(0.013)	(0.012)
Investor characteristics								
Age = 45	-0.050***	0.050***	-0.001	0.004	0.006	-0.123**	-0.019	-0.015
	(0.013)	(0.013)	(0.011)	(0.015)	(0.017)	(0.061)	(0.013)	(0.012)
Age = 60	-0.096***	0.084***	0.012	0.033**	0.034**	-0.137*	-0.081***	-0.076***
	(0.012)	(0.012)	(0.009)	(0.014)	(0.015)	(0.076)	(0.012)	(0.011)
Risk tolerance = high	0.301***	-0.289***	-0.044***	-0.007	0.005	1.284***	0.281***	0.102***
	(0.011)	(0.010)	(0.008)	(0.012)	(0.013)	(0.059)	(0.010)	(0.010)
Home country = US	-0.009	0.009	-0.006	-0.010		0.013	-0.004	-0.013
	(0.014)	(0.013)	(0.006)	(0.015)		(0.060)	(0.013)	(0.011)
Home country = Germany	-0.016	0.019	0.033***	-0.061***	-0.698***	0.017	-0.009	0.011
	(0.019)	(0.017)	(0.010)	(0.020)	(0.015)	(0.116)	(0.018)	(0.016)
Home country = China	0.020	-0.030*	-0.244***	-0.167***	-0.481***	0.317***	-0.068***	-0.170***
	(0.018)	(0.017)	(0.019)	(0.022)	(0.021)	(0.119)	(0.017)	(0.019)
Sustainability = Yes	0.093***	-0.083***	0.003	0.006	-0.019	0.266***	0.023**	-0.000
	(0.012)	(0.011)	(0.009)	(0.013)	(0.015)	(0.070)	(0.011)	(0.011)
Data available	-0.067	0.082**	0.279***	0.464***	-0.358***	0.236	0.150***	0.206***
	(0.042)	(0.039)	(0.040)	(0.039)	(0.044)	(0.227)	(0.035)	(0.038)
Some response error	-0.034***	0.002	-0.023**	0.005	0.084***	-0.219***	-0.077***	-0.063***
	(0.012)	(0.012)	(0.010)	(0.013)	(0.015)	(0.069)	(0.011)	(0.011)
Constant	0.650***	0.297***	0.664***	0.233***	1.041***	2.258***	0.338***	0.619***
	(0.047)	(0.043)	(0.041)	(0.045)	(0.049)	(0.242)	(0.040)	(0.042)
Obs.	1,561	1,561	1,561	1,561	1,264	1,559	1,559	1,559
Adj. R2	0.380	0.382	0.366	0.221	0.596	0.249	0.392	0.241

Note: Alternative assets include commodities, private equity, private debt, cryptocurrencies and real estate (including REITs). Developed markets follow the MSCI definition plus Luxembourg and Liechtenstein. Idiosyncratic volatility is computed as the portfolio-level standard deviation of residuals from FF6 regressions. Omitted categories are "fine-tuned" for LLM type, "$\leq$ 60B" for LLM size, "proprietary" for license type, "30" for age, "low" for risk tolerance, "not specified" for home country, and "no" for sustainability. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$).

Table A16: Risk exposure determinants (robustness: region-specific factor portfolios)

	Monthly rebalancing			Annual rebalancing		
	(1)	(2)	(3)	(4)	(5)	(6)
	Monthly volatility (%)	FF6 β_{mkt}	Idiosyncratic volatility	Monthly volatility (%)	FF6 β_{mkt}	Idiosyncratic volatility
Model characteristics						
Foundation model	-0.079	-0.001	-0.018*	-0.018	-0.001	-0.024**
	(0.077)	(0.011)	(0.010)	(0.130)	(0.011)	(0.010)
Size > 60B	-0.057	0.021**	0.031***	-0.070	0.033***	0.041***
	(0.058)	(0.010)	(0.009)	(0.083)	(0.010)	(0.009)
Open-source	-0.142*	-0.014	-0.020*	-0.079	-0.000	-0.014
	(0.077)	(0.012)	(0.011)	(0.117)	(0.012)	(0.010)
Investor characteristics						
Age = 45	-0.130**	-0.022*	-0.014	-0.105	-0.016	-0.015
	(0.061)	(0.012)	(0.010)	(0.067)	(0.012)	(0.010)
Age = 60	-0.143*	-0.081***	-0.082***	0.027	-0.062***	-0.072***
	(0.075)	(0.012)	(0.010)	(0.120)	(0.011)	(0.010)
Risk tolerance = high	1.287***	0.280***	0.106***	1.079***	0.215***	0.065***
	(0.059)	(0.010)	(0.009)	(0.087)	(0.009)	(0.009)
Home country = US	0.025	0.002	-0.009	-0.027	-0.015	-0.007
	(0.061)	(0.014)	(0.011)	(0.063)	(0.013)	(0.011)
Home country = Germany	-0.064	-0.052***	-0.041***	-0.044	-0.047***	-0.044***
	(0.100)	(0.016)	(0.014)	(0.088)	(0.016)	(0.014)
Home country = China	0.266**	-0.121***	-0.259***	0.548**	-0.101***	-0.232***
	(0.115)	(0.016)	(0.013)	(0.224)	(0.015)	(0.013)
Sustainability = Yes	0.242***	0.018	-0.011	0.348***	0.022**	-0.020**
	(0.070)	(0.011)	(0.009)	(0.108)	(0.010)	(0.009)
Constant	2.439***	0.471***	0.816***	2.359***	0.472***	0.817***
	(0.111)	(0.020)	(0.017)	(0.199)	(0.019)	(0.017)
Obs.	1,559	1,559	1,559	1,559	1,559	1,559
Adj. R2	0.242	0.379	0.314	0.103	0.282	0.251

Note: The table reports coefficients of regressions of risk exposure measures on LLM and investor characteristics (cf. Table 5, columns 7 through 9). The measures are obtained from six-factor regressions using region-specific factor portfolios, i.e., developed market factor portfolios for German and US investor profiles, and emerging market factor portfolios for Chinese investor profiles. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$).

Table A17: Dominance statistics for exposure regressions (LLM recommendations)

	(1) Equity share			(2) Monthly volatility (%)			(3) FF6 β_{mkt}			(4) Idiosyncratic volatility		
	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI
Model characteristics												
Foundation model	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]
Size > 60B	0.01	0.01	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.01]	0.00	0.01	[0.00, 0.01]
Open-source	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]
Investor characteristics												
Age = 45	0.00	0.01	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.00]
Age = 60	0.02	0.06	[0.01, 0.04]	0.00	0.01	[0.00, 0.01]	0.02	0.06	[0.01, 0.03]	0.02	0.12	[0.01, 0.04]
Risk tolerance = high	0.32	0.84	[0.28, 0.35]	0.23	0.92	[0.16, 0.30]	0.32	0.88	[0.28, 0.36]	0.06	0.30	[0.04, 0.08]
Home country = US	0.00	0.01	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]	0.01	0.07	[0.01, 0.02]
Home country = Germany	0.00	0.00	[0.00, 0.01]	0.00	0.00	[0.00, 0.01]	0.00	0.01	[0.00, 0.01]	0.00	0.02	[0.00, 0.00]
Home country = China	0.00	0.01	[0.00, 0.01]	0.01	0.02	[0.00, 0.02]	0.01	0.04	[0.01, 0.02]	0.09	0.47	[0.06, 0.13]
Sustainability = yes	0.03	0.07	[0.01, 0.04]	0.01	0.03	[0.00, 0.02]	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]
Obs.		1,564			1,559			1,559			1,559	
R2		0.38			0.25			0.36			0.19	

Note: The table reports standardized dominance statistics for regressions of equity share (column 1), volatility (column 2), six-factor market beta (column 3), and idiosyncratic volatility (column 4) on the various independent variables. Confidence intervals are based on 10,000 bootstrap samples.

Table A18: Dominance statistics for exposure regressions (robustness: region-specific factor portfolios)

	Monthly rebalancing						Annual rebalancing					
	(1) FF6 β_{mkt}			(2) Idiosyncratic volatility			(3) FF6 β_{mkt}			(4) Idiosyncratic volatility		
	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI
Model characteristics												
Foundation model	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]
Size ¿ 60B	0.01	0.04	[0.00, 0.01]	0.01	0.02	[0.00, 0.01]	0.01	0.02	[0.00, 0.01]	0.01	0.06	[0.01, 0.02]
Open-source	0.00	0.00	[0.00, 0.01]	0.00	0.01	[0.00, 0.01]	0.00	0.00	[0.00, 0.01]	0.00	0.01	[0.00, 0.01]
Investor characteristics												
Age = 45	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]
Age = 60	0.02	0.05	[0.01, 0.03]	0.03	0.09	[0.02, 0.05]	0.02	0.05	[0.00, 0.03]	0.02	0.09	[0.01, 0.04]
Risk tolerance = high	0.33	0.85	[0.29, 0.36]	0.07	0.21	[0.05, 0.09]	0.24	0.84	[0.20, 0.28]	0.03	0.11	[0.01, 0.04]
Home country = US	0.01	0.02	[0.00, 0.01]	0.02	0.08	[0.02, 0.04]	0.00	0.01	[0.00, 0.00]	0.02	0.09	[0.02, 0.04]
Home country = Germany	0.00	0.01	[0.00, 0.01]	0.01	0.02	[0.00, 0.01]	0.00	0.01	[0.00, 0.01]	0.01	0.02	[0.00, 0.01]
Home country = China	0.02	0.06	[0.01, 0.04]	0.18	0.56	[0.15, 0.21]	0.02	0.06	[0.01, 0.03]	0.16	0.61	[0.12, 0.19]
Sustainability = yes	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]	0.00	0.01	[0.00, 0.01]	0.01	0.02	[0.00, 0.01]
Obs.		1,559			1,559			1,559			1,559	
R2		0.38			0.32			0.29			0.26	

Note: The table reports standardized dominance statistics for regressions of six-factor market beta (columns 1 and 3) and idiosyncratic volatility (columns 2 and 4) on the various independent variables. The measures are obtained from six-factor regressions using region-specific factor portfolios, i.e., developed market factor portfolios for German and US investor profiles, and emerging market factor portfolios for Chinese investor profiles.

Table A19: Dominance statistics for exposure regressions (LLM recommendations, robustness: continuous size specification)

	(1) Equity share			(2) Monthly volatility (%)			(3) FF6 β_{mkt}			(4) Idiosyncratic volatility		
	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI	Relative weight	Stand. relative weight	Bootstrapped 95% CI
Model characteristics												
Foundation model	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]
ln (size)	0.00	0.01	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]
Open-source	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]
Investor characteristics												
Age = 45	0.00	0.01	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.00]
Age = 60	0.02	0.06	[0.01, 0.04]	0.00	0.01	[0.00, 0.01]	0.02	0.06	[0.01, 0.03]	0.02	0.12	[0.01, 0.04]
Risk tolerance = high	0.32	0.84	[0.28, 0.35]	0.23	0.92	[0.16, 0.30]	0.32	0.88	[0.28, 0.35]	0.06	0.30	[0.04, 0.08]
Home country = US	0.00	0.01	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]	0.01	0.07	[0.01, 0.01]
Home country = Germany	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.01]	0.00	0.01	[0.00, 0.01]	0.00	0.02	[0.00, 0.00]
Home country = China	0.00	0.01	[0.00, 0.01]	0.01	0.02	[0.00, 0.02]	0.01	0.04	[0.01, 0.02]	0.09	0.47	[0.06, 0.13]
Sustainability = yes	0.03	0.07	[0.01, 0.04]	0.01	0.03	[0.00, 0.02]	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]
Obs.		1,564			1,559			1,559			1,559	
R2		0.38			0.25			0.36			0.19	

Note: The table reports standardized dominance statistics for regressions of equity share (column 1), volatility (column 2), six-factor market beta (column 3), and idiosyncratic volatility (column 4) on the various independent variables. Confidence intervals are based on 10,000 bootstrap samples.

Table A20: Sensitivity of exposure to risk tolerance, by LLM type and size (robustness: region-specific factor portfolios)

	Monthly rebalancing				Annual rebalancing			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	FF6	FF6	Idiosyncratic	Idiosyncratic	FF6	FF6	Idiosyncratic	Idiosyncratic
	β_{mkt}	β_{mkt}	volatility	volatility	β_{mkt}	β_{mkt}	volatility	volatility
Risk tolerance = high	0.238***	0.250***	0.072***	0.091***	0.179***	0.186***	0.038***	0.053***
	(0.015)	(0.014)	(0.013)	(0.013)	(0.014)	(0.014)	(0.013)	(0.014)
Model type								
Risk tolerance = high × original model	0.069***		0.057***		0.058***		0.045***	
	(0.020)		(0.017)		(0.019)		(0.017)	
Model size								
Risk tolerance = high × Size > 60B		0.059***		0.030*		0.055***		0.023
		(0.020)		(0.017)		(0.019)		(0.017)
Constant	0.492***	0.484***	0.833***	0.822***	0.490***	0.484***	0.830***	0.822***
	(0.021)	(0.020)	(0.018)	(0.017)	(0.020)	(0.020)	(0.018)	(0.017)
Obs.	1,559	1,559	1,559	1,559	1,559	1,559	1,559	1,559
Adj. R2	0.384	0.382	0.318	0.315	0.285	0.285	0.254	0.251
Model controls	✓	✓	✓	✓	✓	✓	✓	✓
Profile controls	✓	✓	✓	✓	✓	✓	✓	✓

Note: Model controls include a dummy variable indicating an foundation model (vs. fine-tuned), a dummy variable indicating an open source-LLM (vs. proprietary), and a dummy variable indicating LLM size (> 60B). Profile controls include the investor characteristics listed in Table 5. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$).

Table A21: Sensitivity of exposure to risk tolerance, by LLM type and size (robustness: continuous size specification)

	(1) Equity share	(2) Equity share	(3) Monthly volatility (%)	(4) Monthly volatility (%)	(5) FF6 β_{mkt}	(6) FF6 β_{mkt}	(7) Idiosyncratic volatility	(8) Idiosyncratic volatility
Risk tolerance = high	0.257***	0.139***	1.104***	0.685***	0.244***	0.162***	0.074***	0.056*
	(0.015)	(0.029)	(0.108)	(0.189)	(0.015)	(0.030)	(0.015)	(0.031)
Model type								
Risk tolerance = high × foundation model	0.071***		0.303**		0.065***		0.052***	
	(0.021)		(0.128)		(0.020)		(0.020)	
Model size								
Risk tolerance = high × ln(size)		0.041***		0.152***		0.031***		0.012*
		(0.007)		(0.043)		(0.007)		(0.007)
Constant	0.662***	0.720***	2.741***	2.946***	0.488***	0.528***	0.809***	0.817***
	(0.034)	(0.037)	(0.182)	(0.183)	(0.032)	(0.033)	(0.032)	(0.034)
Obs.	1,564	1,564	1,559	1,559	1,559	1,559	1,559	1,559
Adj. R2	0.378	0.385	0.246	0.249	0.359	0.363	0.191	0.189
Model controls	✓	✓	✓	✓	✓	✓	✓	✓
Profile controls	✓	✓	✓	✓	✓	✓	✓	✓

Note: Developed markets follow the MSCI definition plus Luxembourg and Liechtenstein. Model controls include a dummy variable indicating an foundation model (vs. fine-tuned), an open-source dummy, and the size parameter of each LLM. Profile controls include the investor characteristics listed in Table 5. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$).

Table A22: Performance comparison (LLMs vs. robo-advisors, full regression output)

$\mathbb{1}(LLM)$	Excess return (%)			Sharpe ratio			FF6 α (%)		
	N	Coef.	Adj. R^2	N	Coef.	Adj. R^2	N	Coef.	Adj. R^2
Full sample	1,078	0.159*** (0.014)	0.327	1,078	0.107*** (0.016)	0.244	1,074	-0.046*** (0.010)	0.055
Sub-samples									
Risk tolerance = high	546	0.172*** (0.022)	0.119	546	0.060*** (0.020)	0.055	542	-0.034** (0.016)	0.001
Risk tolerance = low	532	0.145*** (0.015)	0.195	532	0.156*** (0.025)	0.114	532	-0.058*** (0.012)	0.070
Sustainability preference = yes	534	0.178*** (0.019)	0.322	534	0.145*** (0.022)	0.273	530	-0.038*** (0.014)	0.053
Sustainability preference = no	544	0.140*** (0.019)	0.330	544	0.070*** (0.023)	0.215	544	-0.056*** (0.015)	0.051
Age = 30	357	0.137*** (0.024)	0.266	357	0.098*** (0.026)	0.225	355	-0.056*** (0.016)	0.037
Age = 45	362	0.144*** (0.024)	0.330	362	0.089*** (0.028)	0.220	360	-0.056*** (0.016)	0.052
Age = 60	359	0.194*** (0.022)	0.360	359	0.134*** (0.028)	0.257	359	-0.032* (0.019)	0.075
Home country = United States	551	0.205*** (0.019)	0.364	551	0.175*** (0.023)	0.271	547	-0.017 (0.014)	0.016
Home country = Germany	527	0.110*** (0.019)	0.274	527	0.036* (0.021)	0.210	527	-0.075*** (0.014)	0.101
Experience = high (robo-advisors)	838	0.144*** (0.016)	0.295	838	0.091*** (0.020)	0.214	836	-0.051*** (0.011)	0.052
Experience = low (robo-advisors)	838	0.173*** (0.016)	0.303	838	0.123*** (0.020)	0.232	836	-0.042*** (0.012)	0.051

Note: The table reports the full regression outputs for regressions of portfolio performance on a dummy variable indicating that a portfolio was recommended by an LLM (rather than a robo-advisor) and profile characteristics. We control for all investor characteristics in the full sample and all investor characteristics except the one identifying the respective sub-sample in the sub-samples. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$).

Table A23: Determinants of portfolio performance and fees (robustness: implementability)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Excess return (%)	Excess return (%)	Sharpe ratio	Sharpe ratio	FF6 α	FF6 α	Total exp. ratio (%)	Total exp. ratio (%)
Model characteristics								
Foundation model	0.008	0.004	-0.016	-0.032	0.017	0.023	-0.008	0.015
	(0.019)	(0.021)	(0.017)	(0.021)	(0.015)	(0.016)	(0.011)	(0.015)
Size > 60B	-0.000	-0.011	-0.014	-0.033**	-0.012	-0.010	-0.023**	0.002
	(0.014)	(0.016)	(0.015)	(0.016)	(0.010)	(0.012)	(0.009)	(0.009)
Open-source	-0.041**	-0.006	-0.028	0.001	-0.057***	-0.023	-0.021*	0.011
	(0.020)	(0.024)	(0.019)	(0.021)	(0.016)	(0.020)	(0.011)	(0.012)
Cutoff within 6 months		-0.005		-0.048**		0.019		0.040***
		(0.018)		(0.020)		(0.013)		(0.013)
Data available	0.429***	0.448***	0.511***	0.530***	0.126***	0.118***	-0.247***	-0.191***
	(0.042)	(0.051)	(0.049)	(0.056)	(0.033)	(0.043)	(0.060)	(0.068)
Some response error	-0.019	-0.015	-0.022	-0.010	0.034**	0.032**	0.057***	0.046***
	(0.018)	(0.020)	(0.017)	(0.018)	(0.014)	(0.015)	(0.010)	(0.011)
Constant	-0.130***	-0.169***	-0.127**	-0.129**	-0.256***	-0.281***	0.373***	0.243***
	(0.049)	(0.063)	(0.053)	(0.064)	(0.036)	(0.051)	(0.062)	(0.065)
Obs.	1,559	1,186	1,559	1,186	1,559	1,186	1,534	1,166
Adj. R2	0.266	0.286	0.249	0.265	0.069	0.060	0.158	0.125
Profile controls	✓	✓	✓	✓	✓	✓	✓	✓

Note: Omitted categories are "fine-tuned" for LLM type, "$\leq$ 60B" for LLM size, "proprietary" for license type, "30" for age, "low" for risk tolerance, "not specified" for home country, and "no" for sustainability. Profile controls include the investor characteristics listed in Table 9. Heteroskedasticity-robust standard errors are reported in parentheses $({}^{*}p < 0.1, {}^{**}p < 0.05, {}^{***}p < 0.01)$.

Table A24: Determinants of portfolio performance (robustness: region-specific factor portfolios)

	Monthly rebalancing		Annual rebalancing	
	(1)	(2)	(3)	(4)
	FF6	FF6	FF6	FF6
	α	α	α	α
Model characteristics				
Foundation model	0.005	0.006	-0.009	0.006
	(0.017)	(0.017)	(0.022)	(0.021)
Size > 60B	-0.014	-0.016	-0.015	-0.009
	(0.010)	(0.013)	(0.011)	(0.014)
Open-source	-0.045**	-0.019	-0.044*	-0.027
	(0.018)	(0.024)	(0.023)	(0.032)
Cutoff within 6 months		0.010		0.031**
		(0.012)		(0.014)
Investor characteristics				
Age = 45	-0.015	-0.006	0.002	0.012
	(0.013)	(0.014)	(0.014)	(0.016)
Age = 60	0.018	0.024	0.050***	0.057***
	(0.015)	(0.017)	(0.017)	(0.020)
Risk tolerance = high	0.089***	0.078***	0.052***	0.041**
	(0.012)	(0.013)	(0.014)	(0.016)
Home country = US	0.019*	0.012	0.024*	0.014
	(0.012)	(0.013)	(0.013)	(0.015)
Home country = Germany	-0.045***	-0.063***	-0.016	-0.038**
	(0.015)	(0.016)	(0.017)	(0.018)
Home country = China	0.226***	0.206***	0.262***	0.248***
	(0.025)	(0.029)	(0.030)	(0.037)
Sustainability = Yes	0.003	-0.008	0.021	0.016
	(0.013)	(0.014)	(0.015)	(0.018)
Constant	-0.119***	-0.135***	-0.160***	-0.198***
	(0.021)	(0.027)	(0.026)	(0.034)
Obs.	1,559	1,186	1,559	1,186
Adj. R2	0.153	0.143	0.125	0.119

Note: Omitted categories are "fine-tuned" for LLM type, "$\leq$ 60B" for LLM size, "proprietary" for license type, "30" for age, "low" for risk tolerance, "not specified" for home country, and "no" for sustainability. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1$,** $p < 0.05$,*** $p < 0.01$).

Table A25: Home bias determinants

	All countries			China	Germany	United States
	(1)	(2)	(3)	(4)	(5)	(6)
	$d_i - m_i$	$\frac{d_i - m_i}{m_i}$	$\frac{d_i - m_i}{1 - m_i}$	$d_i - m_i$	$d_i - m_i$	$d_i - m_i$
Model characteristics						
Foundation model	0.007	-0.097	0.022	-0.018	-0.006	0.018
	(0.015)	(0.436)	(0.032)	(0.040)	(0.022)	(0.021)
Size > 60B	0.044***	0.934**	0.083***	0.062	0.030	0.043**
	(0.015)	(0.457)	(0.031)	(0.041)	(0.021)	(0.020)
Open-source	-0.045***	-1.705***	-0.057	-0.152***	-0.034	-0.010
	(0.017)	(0.515)	(0.036)	(0.048)	(0.025)	(0.023)
Profile characteristics						
Age = 45	0.005	0.116	0.011	-0.032	0.032	-0.004
	(0.018)	(0.610)	(0.035)	(0.044)	(0.025)	(0.028)
Age = 60	0.032**	0.295	0.079**	-0.022	0.045*	0.045**
	(0.016)	(0.452)	(0.033)	(0.046)	(0.025)	(0.019)
Risk tolerance = high	-0.020	-0.517	-0.037	0.024	-0.065***	-0.019
	(0.014)	(0.424)	(0.028)	(0.037)	(0.022)	(0.018)
Home country = Germany	-0.030**	3.500***	-0.202***			
	(0.015)	(0.594)	(0.028)			
Home country = China	0.113***	8.846***	-0.053*			
	(0.021)	(0.802)	(0.032)			
Sustainability preference = Yes	-0.013	-0.894*	-0.005	-0.018	-0.057***	0.016
	(0.015)	(0.522)	(0.028)	(0.037)	(0.021)	(0.022)
Constant	0.106***	1.196*	0.244***	0.307***	0.115***	0.065*
	(0.025)	(0.675)	(0.056)	(0.064)	(0.042)	(0.034)
Obs.	1,230	1,230	1,230	271	277	682
Adj. R2	0.062	0.188	0.041	0.051	0.063	0.014

Note: The table reports the coefficients of regressions of various home bias measures on LLM and profile characteristics. Columns 1 through 3 pool observations from all three home countries. Columns 4 through 6 use data only for Chinese, German, and US profiles, respectively. Heteroskedasticity-robust standard errors are reported in parentheses (${}^{*}p < 0.1,{}^{**} p < 0.05,{}^{***} p < 0.01$).

Table A26: Domestic allocation and home bias (robo-advisors)

	All asset classes			Equity only			Home bias				
Home Country	Mean	N	SD	d_i	N	SD	m_i	$d_i - m_i$	t ($d_i = m_i$)	$\frac{d_i - m_i}{m_i}$	$\frac{d_i - m_i}{1 - m_i}$
Panel A: All Observations											
Germany	0.05	240	0.03	0.03	240	0.01	0.02	0.01	13.73***	0.43	0.01
United States	0.70	240	0.12	0.59	240	0.16	0.63	-0.04	-3.57***	-0.06	-0.10
Panel B: Only high risk-tolerance profiles											
Germany	0.04	120	0.03	0.03	120	0.01	0.02	0.01	9.71***	0.44	0.01
United States	0.66	120	0.10	0.62	120	0.08	0.63	-0.01	-1.02	-0.01	-0.02
Panel C: Only high experience profiles											
Germany	0.05	120	0.03	0.03	120	0.01	0.02	0.01	9.64***	0.43	0.01
United States	0.70	120	0.12	0.59	120	0.16	0.63	-0.04	-2.51***	-0.06	-0.10

Note: The table reports the portfolio weights assigned to domestic securities by home country of the investor profile. The first column reports the average domestic allocation d_i, the number of portfolios, and the standard deviation of the domestic allocation within the equity portion of a portfolio. The second column reports the respective country weight m_i in the market portfolio as proxied by the MSCI All-Country World Investable Markets Index, as well as three home bias measures (simple weight-gap, weight-gap divided by benchmark weight, weight-gap divided benchmark international weight, cf. Cooper et al., 2018). The table further reports the test statistics of one-sample t-tests for equality of the domestic equity weight d_i and the benchmark weight m_i. Significance is indicated by stars ($^{*}p < 0.1,^{**} p < 0.05,^{***} p <$ 0.01). Panel A uses all observations. Panel B (C) only uses portfolio recommendations for high-risk-tolerance (high experience) investor profiles.

Table A27: Effect of home bias correction prompt

	All recommendations	Only recommendations without errors
	(1)	(2)
	Domestic equity share	Domestic equity share
Correction	-0.153***	-0.163***
	(0.024)	(0.029)
Obs.	450	300
Adj. R2	0.499	0.503
Model controls	✓	✓
Profile controls	✓	✓

Note: The table reports the coefficients of OLS regressions of domestic exposure on a dummy variable indicating whether a correction prompt for the respective error has been included (Correction), as well as LLM and profile characteristics. For column 1, we use all recommendation pairs without errors in the equity portion of the portfolio recommendations. For column 2, we only use recommendation pairs without any response errors. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$).

Table A28: Portfolio exposures by gender

		(1)		(2)		(1) - (2)	
		Male		Female			
		Mean	N	Mean	N	Δ	z score
Panel A: Equity share							
Risk tolerance = high	Age = 30	0.84	24	0.83	26	0.01	0.25
Risk tolerance = high	Age = 60	0.85	24	0.83	24	0.02	0.32
Risk tolerance = low	Age = 30	0.56	25	0.53	22	0.03	0.47
Risk tolerance = low	Age = 60	0.40	23	0.36	27	0.04	0.70
Panel B: FF6 market beta							
Risk tolerance = high	Age = 30	0.77	24	0.74	26	0.03	0.91
Risk tolerance = high	Age = 60	0.79	24	0.76	24	0.03	0.36
Risk tolerance = low	Age = 30	0.48	25	0.47	22	0.01	0.27
Risk tolerance = low	Age = 60	0.35	23	0.30	27	0.05	0.79
Panel C: Idiosyncratic volatility							
Risk tolerance = high	Age = 30	0.91	24	0.89	26	0.02	0.89
Risk tolerance = high	Age = 60	0.90	24	0.88	24	0.02	-0.30
Risk tolerance = low	Age = 30	0.80	25	0.83	22	-0.03	0.37
Risk tolerance = low	Age = 60	0.70	23	0.65	27	0.05	0.50
Panel D: Domestic share							
Risk tolerance = high	Age = 30	0.72	24	0.67	26	0.05	0.74
Risk tolerance = high	Age = 60	0.71	24	0.73	24	-0.02	-0.09
Risk tolerance = low	Age = 30	0.72	25	0.75	22	-0.03	-0.23
Risk tolerance = low	Age = 60	0.74	23	0.73	27	0.01	0.07
Panel E: Individual stock share							
Risk tolerance = high	Age = 30	0.02	32	0.04	32	-0.02	-0.49
Risk tolerance = high	Age = 60	0.05	32	0.06	32	-0.01	0.33
Risk tolerance = low	Age = 30	0.01	32	0.00	32	0.01	1.00
Risk tolerance = low	Age = 60	0.01	32	0.01	32	0.00	-0.02

Note: The table reports the average equity share (panel A), six-factor market beta (panel B), idiosyncratic volatility (panel C), share of domestic securities (panel D), and share of individual stocks (panel E) of portfolios recommended to male (column 1) and female (column 2) investors, separately by risk tolerance and age. Column 3 reports the difference in means and the z score of non-parametric rank tests for differences in means ($^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$).

Table A29: Heterogeneity in gender-based discrimination

	(1) Δ equity share	(2) Δ developed markets	(3) Δ FF6 β_{mkt}	(4) Δ IVOL	(5) Δ ind. stocks
Model characteristics					
Foundation model	0.031	-0.009	0.046	0.021	-0.026
	(0.052)	(0.021)	(0.041)	(0.034)	(0.022)
Size > 60B	-0.046	0.012	-0.026	0.015	0.023
	(0.044)	(0.020)	(0.036)	(0.037)	(0.015)
Open-source	0.016	0.003	0.025	0.100**	-0.021
	(0.049)	(0.022)	(0.045)	(0.041)	(0.020)
Constant	0.027	0.001	-0.004	-0.065	0.014
	(0.071)	(0.033)	(0.056)	(0.051)	(0.018)
Obs.	88	88	88	88	128
Adj. R2	-0.014	-0.025	-0.010	0.041	0.005

Note: The table reports the coefficients of regressions of gender-based differences in various exposure measures on LLM characteristics. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$).

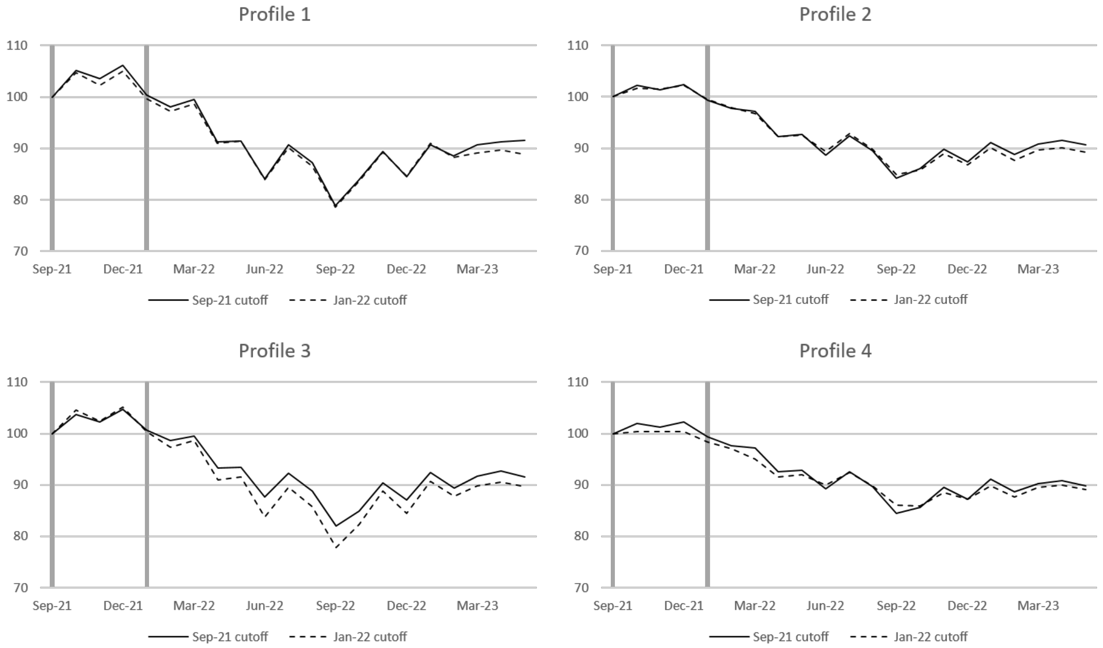
Table A30: LLM leaguetable

		Implementability		**Exposure**		**Performance**		
Rank	**Model**	**Data available**	**Some resp. error**	**Home bias**	**Δ equity share**	**Sharpe ratio**	**FF6 α (% p.m.)**	**Total exp. ratio (%)**
1	Baichuan2	0.99	0.14	0.09	0.51	0.23	-0.06	0.09
2	Solar	1.00	0.31	-0.06	0.31	0.28	-0.01	0.12
3	alpaca-7b	1.00	0.58	-0.03		0.22	-0.09	0.07
4	Deepseek	0.99	0.09	0.07	0.48	0.16	-0.13	0.11
5	Yi-34B-Chat	0.98	0.20	0.06	0.39	0.19	-0.15	0.08
6	Jurassic-2 Ultra	0.99	0.23	0.15	-0.02	0.43	0.18	0.16
7	PPLX-70B	0.98	0.41	0.13	0.28	0.21	0.01	0.13
8	falcon-40b-instruct	1.00	0.98	0.10	0.31	0.23	0.01	0.27
9	llama-2-70b-chat	0.98	0.63	-0.03	0.34	0.16	-0.17	0.14
10	Platypus2-70B-instruct	0.98	0.59	0.02	0.22	0.19	-0.13	0.14
11	OpenHermes-2p5-Mistral-7B	0.94	0.41	0.05	0.10	0.24	-0.15	0.14
12	GPT-3.5-Turbo	0.91	0.31	0.08	0.27	0.20	-0.16	0.13
13	GPT-4-Turbo	0.87	0.14	0.13	0.44	0.12	-0.14	0.16
14	Gemini-Pro	0.89	0.19	0.22	0.39	0.15	-0.09	0.16
15	openchat-3.5-1210	0.98	0.41	0.01	0.23	0.17	-0.20	0.16
16	chatglm3-6b	0.93	0.92	0.21	0.15	0.26	0.04	0.14
17	Tulu-2-DPO	0.93	0.13	0.12	0.40	0.10	-0.18	0.19
18	mixtral-8x7b-instruct-v0.	0.92	0.19	0.11	0.32	0.14	-0.17	0.18
19	orca-2-13b	0.99	0.75	0.00	-0.03	0.15	-0.24	0.13
20	LLaMA-20-70B-SteerLM-Chat	0.91	0.45	0.22	0.34	0.18	-0.04	0.23
21	Qwen-72B-Chat	0.88	0.44	0.13	0.42	0.08	-0.18	0.15
22	Claude 2.1	0.83	0.28	0.16	0.24	0.16	-0.08	0.24
23	RedPajama-INCITE-7B-Chat	0.94	0.95	0.10	0.17	0.17	-0.13	0.30
24	starling-lm-7b-alpha	0.99	0.78	0.21	0.25	0.14	-0.14	0.23
25	zephyr-7b-beta	0.95	0.38	0.09	0.21	0.09	-0.20	0.24
26	Mistral-Medium	0.84	0.22	0.22	0.28	0.15	-0.15	0.24
27	WizardLM	0.89	0.55	0.17	0.43	0.11	-0.19	0.19
28	Command-Nightly	0.93	0.61	0.23	0.34	-0.02	-0.04	0.36
29	Claude 2.0	0.87	0.27	0.17	0.29	0.05	-0.18	0.21
30	Bard-Jan-24-Gemini-Pro	0.79	0.63	0.40	0.53	-0.02	-0.13	0.24
31	vicuna-33b-v1.3		1.00					
32	StripedHyena-Hessian-7B		1.00					
	Average	**0.93**	**0.47**	**0.12**	**0.30**	**0.16**	**-0.11**	**0.18**

Note: The table reports average suitability measures by LLM. Home bias is computed as the absolute difference of domestic weights from the respective country's weight in the MSCI All-Country World Index (63% US, 3% Germany, 2% China). Δ equity share is the difference between the average equity share for high-risk-tolerance profiles and low-risk-tolerance risk profiles. The ranking is computed as the un-weighted average of all seven sub-rankings.

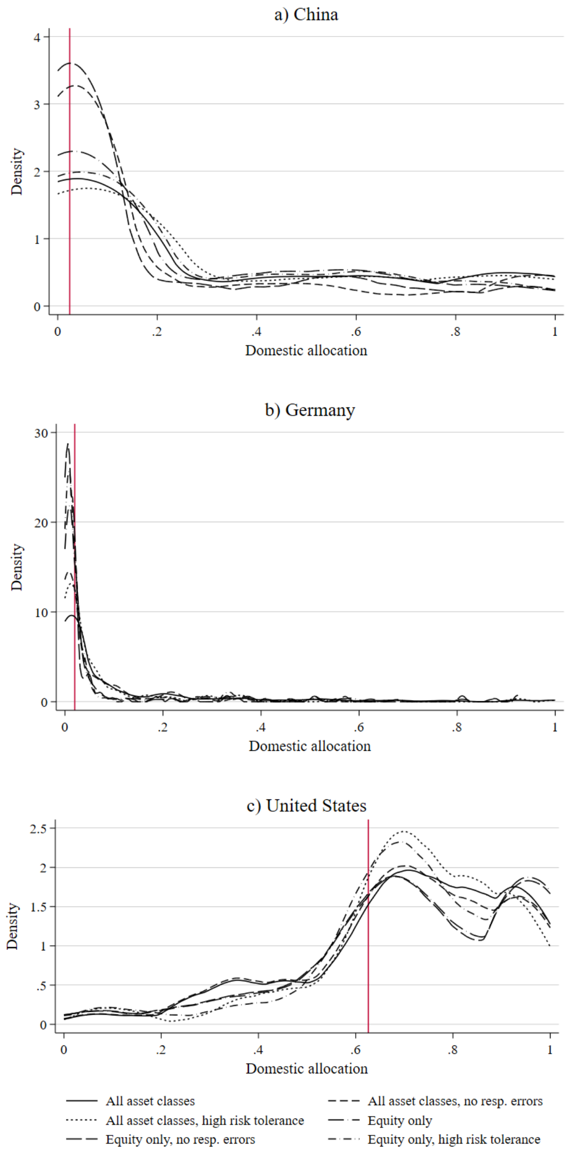
2. Figures

Figure A1: Impact of GPT-4 update on portfolio suggestions



Note: This figure shows the evolution of $100 invested in each of the portfolios from Sep 30, 2021 to May 31, 2022. The solid line shows the evolution of the originally suggested portfolios (elicited May 2023), which are based on data up until Sep 2021. The dashed line shows the evolution of the updated suggestions (elicited Oct 2023), which are based on data up until Jan 2022. Vertical bars represent the information cut-off dates of the initial and updated recommendations (Sep 21 and Jan 22).

Figure A2: Domestic allocation, by home country



Note: The figures display density plots of the portfolio weights assigned to domestic securities by home country of the investor profile. The vertical red lines depict the baseline weights m_i. The various lines represent various sub-samples (as listed in Table 10).