Using large language models for financial advice*

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

September 10, 2024

Abstract

We study whether large language models (LLMs) are able to generate suitable financial advice and which model features are associated with higher-quality advice. To this end, we elicit portfolio recommendations from 32 LLMs for 64 investor profiles differing with respect to their risk tolerance and capacity, home country, sustainability preferences, gender, and investment experience. Our results suggest that LLMs are generally capable of generating financial advice as the recommendations can in fact be implemented, take into account important investor characteristics when determining market and risk exposures, and display historical performance in line with assumed risks. We further find that foundation models and larger models generate portfolios that are easier to implement and more sensitive to investors' circumstances than fine-tuned models and smaller models. Some of our results are consistent with LLMs inheriting human biases as domestic exposures substantially exceed benchmark weights (home bias). On the other hand, we find no evidence of gender-based discrimination found in human financial advice. Based on these results, we discuss the potential application of LLMs in the financial advice context.

Keywords: Generative AI · artificial intelligence · large language models · financial advice · portfolio management
JEL-Classification: G00 · G11 · G40

---

*A previous version of this working paper was published under the title "Using GPT-4 for financial advice" (Fieberg et al., 2023). We thank Niklas Töllich for excellent research assistance. We are also grateful to Tom Dudda, Alejandro Lopez Lira, Matthias Mattusch, Vatsala Shreeti (discussant), Michael Weber, as well as participants of the 8th Luxembourg Workshop on Household Finance and Consumption 2023 and the 2024 Journal of Banking & Finance "Generative AI in Finance" paper development workshop in Dresden for their valuable suggestions. All errors remain our own.

†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

Text-based generative artificial intelligence (AI) tools enable tech-savvy laypeople to conduct tasks in complex domains without much prior experience. For example, early users of GPT-4, a pre-trained transformer-based chatbot developed by OpenAI, marvel at the tool's ability to draw on the vast collection of texts it was trained on to draft high-quality term papers and poems or spot mistakes in computer code (The New York Times, 2022; Feng et al., 2023). Dowling and Lucey (2023) assess the use of GPT-4 in producing Finance research and find that it generates plausible research ideas, literature reviews, and test suggestions. Given the remarkable quality of output even in complex and creative applications such as academic research, it stands to reason that service providers relying heavily on freely available information will be rendered obsolete by advanced AI tools. Thus, this study investigates whether large language models (LLMs) are a suitable source of financial advice. Specifically, we assess the quality of portfolio recommendations generated by off-the-shelf LLMs for various investor profiles.

The financial advice domain provides a compelling application of LLMs for two reasons. First, LLMs may accelerate the democratization of access to financial advice, which has traditionally been limited to wealthy investors, through digital technology. Financial advisors are already using rule-based financial advice systems (robo-advisors) to provide low-cost portfolio management services at scale (Jung et al., 2018; Rühr et al., 2019). Because interaction with a human advisor is limited, fees and minimum investment amounts are much lower than for traditional human advice offerings (Philippon, 2016). LLMs may provide substantial improvements in the quality of digital financial advice due to their conversational capacity (cf. Lo and Ross, 2024), which may help further reduce the required human interaction. On the one hand, they are able to elaborate on the rationale behind a specific recommendation, which may increase adoption (Litterscheidt and Streich, 2020). This 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).[1] On the other hand, they are capable of simulating a personal relationship with the client (Safdari et al., 2023), which the financial advice literature considers an important aspect of financial advice provision (Gennaioli et al., 2015). Stolper and Walter (2019), for instance, show that when 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 LLMs may be well-suited to provide financial advice. LLMs such as GPT-4 have been shown to be able to reproduce the correct answers to standard financial literacy questions (Niszczota

[1] Germann and Merkle (2023) find no evidence of general algorithm aversion in an experimental financial decision-making context. However, 44% of participants in their study initially choose to delegate to a human rather than an algorithmic decision support system, which suggests further adoption potential.

and Abbas, 2023b) and 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). As opposed to human investors, who can be overwhelmed by large amounts of financial information (Hirshleifer et al., 2009; Frederickson and Zolotoy, 2016; Schmidt, 2019), LLMs may thus be able to efficiently distill relevant information from financial texts. Furthermore, LLM's 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 sucessfully compute the stocks' market betas and other portfolio metrics. Thus, state-of-the-art models 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 model 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). To assess the ability of LLMs to provide financial advice to international investors, we include Chinese, German, and US investor profiles to reflect large economies from both developed and emerging markets. For each of the 64 profiles, we elicit portfolio recommendations from 32 commonly used LLMs.[2] We categorize models 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 improving reasoning capabilities). While there are models that are fine-tuned for the financial context (Wu et al., 2023), none of the models we consider are fine-tuned to the financial context in general, nor the financial advice context specifically. We use the number of parameters used in the training of a model as a measure of its complexity and size. Larger models have been shown to learn and retain more complex relationships within the training data, which allows them to produce higher-quality

[2] We compile a list of models from common LLM leaderboards. Details on the model selection are provided in section 3.3.

output with higher generalizability than smaller models (Ding et al., 2023; Brown et al., 2020). For models published under an *open-source license*, information on the training process, input data, and model configuration are made available to the developing and research communities. Developers of *proprietary models*, on the other hand, do not publish details on their models.

Based on the results of an emerging range of studies investigating the abilities of LLMs and the required capabilities as inferred from literature from the financial advice domain, we formulate hypotheses regarding the use of LLMs in the financial advice context (section 2). First, we hypothesize that LLMs should in general be able to provide financial advice due to the information contained in commonly used training data and the results of early capability tests. Second, 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. Third, we expect more complex models (as measured by the number of parameters) to be better suited to provide financial advice due to the higher generalizability of output and scaling properties observed in other tasks.

We then investigate three distinct suitability dimensions for each of the resulting 2,048 portfolio recommendations (64 profiles $\times$ 32 models). 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., wrong tickers) 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 more difficult to implement. Second, we are interested in the extent to which the portfolios' exposure to asset classes, markets, and various risk components (total risk, market risk, and idiosyncratic risk) is in line with the investor profiles' characteristics. In line with modern portfolio theory, we focus on risk tolerance as the main determinant of portfolio risk. Third, we are interested in the historical (risk-adjusted) performance of the recommended portfolios.[3] We use three measures of performance: the average monthly excess return, average annual Sharpe ratio, and alphas to a six-factor model. To account for the operating expenses of implementing a portfolio, we additionally include the portfolios' total expense ratio (TER) in our performance analyses.

Our results can be summarized as follows. First, LLMs are generally able to recommend portfolios that (i) can in fact be implemented, (ii) take into account central investor characteristics (most importantly, risk tolerance) in determining suitable exposures to markets, geographies, and risk factors, and (iii) display historical performance which is in line with benchmark figures from the literature and appropriate given their risk level.

Second, foundation models and larger models provide portfolio recommendations which are easier to implement than those of fine-tuned models as the likelihood of response errors

[3] Since our tests suffer from look-ahead bias, the performance results should be interpreted with caution.

is lower. Our results also suggest that foundation models and larger models recommend portfolios that are more suitable for the respective investor characteristics than those of fine-tuned models. Specifically, foundation models and larger models differentiate exposure to equity to a greater extent depending on an investor's risk tolerance. We find no systematic differences in the risk-adjusted returns of foundation versus fine-tuned models or larger versus smaller models.

Third, we find some evidence that LLMs internalize biases from their training data. Specifically, we find that on average, recommendations suffer from home bias, i.e., excessive allocation to domestic securities (Coval and Moskowitz, 1999). We find no evidence of gender-based discrimination in the exposure of the portfolios as was documented in a recent audit study involving human advisors (Bhattacharya et al., 2024).

The remainder of the paper is organized as follows. Section 2 derives testable hypotheses based on related literature on LLM capabilities and requirements for financial advice. Section 3 details the methodology we employ to elicit portfolio recommendations, as well as the construction of 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?

Do we expect LLMs to be able to generate suitable financial advice (RQ1)? We argue that two pre-conditions must be met: LLMs must have access to a certain set of information in their training data and they must have the capacity to interpret this information and generate financial advice based on it. We will elaborate on each of the two pre-conditions in the following.

We argue that LLMs must have access to information in three domains: First, they must have access to basic financial theory including modern portfolio theory and financial advice provision (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. Bhattacharya et al. (2012), who design a financial advice intervention based on mean-variance optimization, include the concepts of diversification, mean-variance efficiency, Sharpe ratio, and tax advantages in their advice documentation.

Second, LLMs must have access to a range of specific financial products as well as some details on their exposure and risk-return profile. Rule-based digital financial advice systems typically draw on a pre-selected set of investment products (mostly exchange-traded funds), whose portfolio weights are determined by an investor's individual circumstances (Rühr et al., 2019). Bhattacharya et al. (2012) include a fact sheet for each security recommended by their financial advice system, which typically includes product identifiers (e.g., ISIN or ticker), the

historical risk-return profile, exposure to regions, industries, and individual securities (for funds), and sustainability characteristics. 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 '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. News affect the market's assessment of the levels and variability of future returns (Cready and Gurun, 2010), which should affect the suggested portfolio strategy. Some empirical studies document a delayed reaction to company news (Chan, 2003), especially when investors are inattentive (Hirshleifer et al., 2009; Drake et al., 2012; Schmidt, 2019), which implies that timely access to news could improve portfolio performance.

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 is one of the most commonly used datasets used for training LLMs (Zhao et al., 2023). CommonCrawl is one of the largest publically available datasets and is compiled through regular scrapes of internet content since 2008. Most models use pre-filtered subsets of the dataset for training purposes. We scan the list of domains contained in the dataset and identify domains which provide information in each of the three blocks discussed above.[4] Table 1 lists domains identified in the CommonCrawl dataset that correspond to the three information blocks. For information on basic financial theory, CommonCrawl contains the domains of reputable educational institutions (such as Yale, Duke, and Berkeley), general-purpose knowledge domains, and even Finance-specific domains. In principle, this affords LLMs access to college-level Finance course content (which is occasionally made publicly available), superficial descriptions of the main Finance theories, and even sophisticated valuation techniques. For specific information on investment products, the dataset contains the domains of platforms that can be used for financial analysis. For instance, TradingView, which is contained in the CommonCrawl data, provides tickers, exposure descriptions, historical price data, and technical indicators on a wide range of individual securities and funds. Finally, the dataset also contains domains providing time-stamped company and market news (e.g., Euronews). Taken together, CommonCrawl, which is part of most models' training data, should contain information on all three information domains deemed necessary to provide suitable financial advice.[5]

Aside from access to information, LLMs must also have the capacity to correctly interpret the contained information to form recommendations based on it. Regarding basic financial

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

[5] Starling 7B alpha, one of the models we investigate, references investopedia.com, a provider of investment-related definitions, concepts, and news, as well as vanguard.com and fidelity.com, two fund providers, as additional resources.

concepts, Niszczota and Abbas (2023a) assess GPT-3.5 and GPT-4's ability to answer a set of 21 single-choice questions commonly used to assess financial literacy. The questions elicit the respondent's knowledge of the relationship between interest rates and bond prices, compound interest, the risk profile of an individual stock versus that of a stock fund, and risk diversification (Mitchell and Lusardi, 2022; Heinberg et al., 2014). The results suggest that advanced models 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 that less than half of US respondents were able to answer the 'Big Three' financial literacy questions (Lusardi and Mitchell, 2023). While these results suggest that (advanced) 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.[6]

Regarding information on specific investment products, we document in a preliminary version of this study that GPT-4 is indeed able to produce suitable portfolio suggestions containing specific tickers and security names when prompted for financial advice (Fieberg et al., 2023). Besides, 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.

Finally, there is 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 daily risk-adjusted stock returns correlate significantly with the classifications. Consistent with limits to arbitrage driving abnormal returns, predictability is stronger for smaller stocks (which display higher trading frictions) and following negative news (for which short-sale constraints may slow price discovery). Aside from classifying financial news, LLMs are also able to synthesize the most important pieces of information from text: Kim et al. (2023a) show that sentiment scores derived from the summaries of corporate disclosures generated by LLMs are superior predictors of subsequent stock price movements than sentiment scores derived from the raw disclosures. The findings of these studies are corroborated by a similarly 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).

Taken together, we generally expect LLMs to be capable of providing financial advice as the relevant information is likely contained in most models' training data and early studies mostly using OpenAI's GPT models document strong capabilities in reproducing basic fi-

[6] In fact, Smith (2024) shows that GPT-4 ignores the time value of money when prompted for advice on a car loan.

Table 1: 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)

nancial concepts, using financial data for portfolio management purposes, and interpreting financial news.

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

2.2 Which model features are associated with better financial advice?

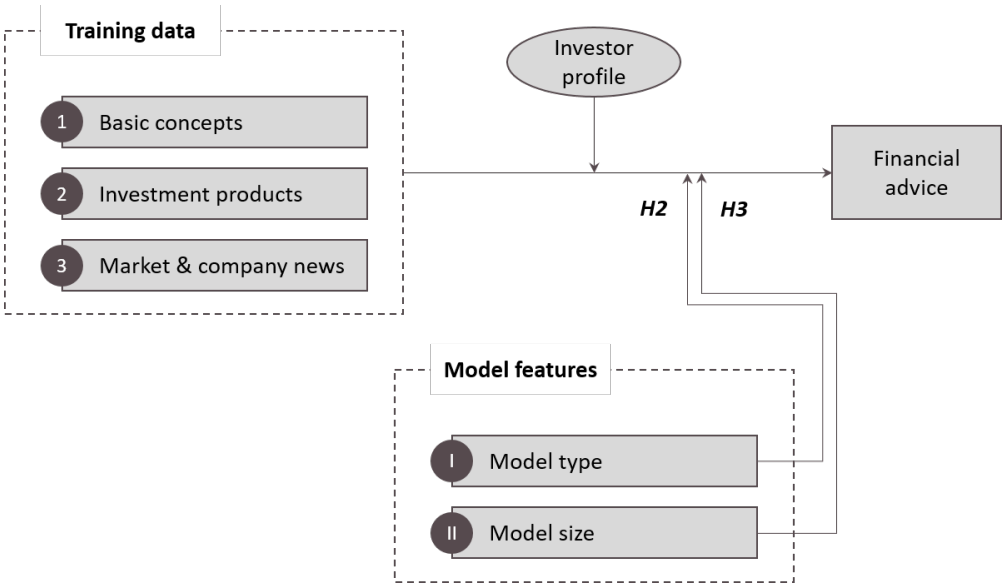
Aside from the general question of whether LLMs are able to provide financial advice, we are also interested in the relationship between specific model features and the quality of financial advice. Within our theoretical framework, model features may affect the quality of generated advice by determining how a set of training data is used to generate financial advice (see Figure 1). Specifically, we will investigate differences in the models' ability to generate financial advice regarding their types (foundation model vs. instruction-fine-tuned model) and complexity (number of parameters).[7]

Model types — 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 on a narrower 'downstream' dataset to suit a specific purpose (e.g., conversational capacity, reasoning capability, forecasting) using machine learning techniques (Ovadia et al., 2023). The fine-tuned models in our sample are mostly re-trained to improve conversational abilities. While fine-tuning can improve a model's performance in a specific task, there are some risks associated with it. This is particularly true if these models are used outside the domains they have been re-trained for. Since none of the models we consider in this analysis are specifically fine-tuned for application in the financial advice context, these risks apply.[8] First, excessive fine-tuning using small downstream datasets can lead to poor model performance outside the intended domain due to overfitting (overly adjusting weights to the narrow dataset) and 'catastrophic forgetting' (omission of previously learned relationships due to parameter updates during the fine-tuning stage, Luo

[7] When investigating model performance, we argue that a model's information cut-off may affect the quality of a model's training data. Models with regular information updates or more recent information cut-offs should have more recent information on financial news and prior returns, which should improve performance.

[8] 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 models narrowly trained on this task (Kim et al., 2024; Li et al., 2023).

Figure 1: Theoretical framework



et al., 2023).[9] Second, fine-tuning may increase the likelihood of hallucination when there is a discrepancy between the original training data and the downstream dataset (capability misalignment) or when models are trained to provide answers that appease human evaluators (belief misalignment, cf. Zhao et al., 2023; Huang et al., 2023). Taken together, we expect foundation models to be better suited to provide financial advice given the risks described above.

Hypothesis 2. *Foundation models are better suited to provide financial advice than fine-tuned models (i.e., models re-trained to improve conversational capabilities).*

Model size — We measure model *complexity* using the number of parameters a model employs. Models 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 models tend to produce higher-quality output with higher generalizability than smaller models, (Ding et al., 2023; Brown et al., 2020). Ding et al. (2023) compare the performance of variously sized models and find relatively smooth scaling properties. They also find that the performance gap is larger for few-shot evaluations than for zero-shot or one-shot evaluations, suggesting that larger models are better able to meta-learn from tasks. In the Finance domain, Lopez-Lira and Tang (2023) demonstrate that greater model complexity enhances the capability to anticipate future stock price movements from news. They show that larger models (GPT-4) outperform smaller models (GPT-1, GPT-2, BERT), especially for complex news articles. While the models do not only differ with respect to their complex-

[9] Approaches to mitigate this include, but are not limited to, 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).

ity, the results suggest that return predictability based on LLM-assessed news headlines is an emerging capacity of larger (i.e., more complex) LLMs. Taken together, we expect larger models to be better suited to provide financial advice.

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

3 Method

This section details our methodological choices in defining investor profiles (section 3.1), formulating the prompt texts used to elicit recommendations (section 3.2), compiling a list of recent, well-performing LLMs suitable for our assessment (section 3.3), and constructing the various suitability measures for our analyses (section 3.4).

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 and risk tolerance, their home country, sustainability preferences, gender and investment experience (see Table B1 in the Appendix).

According to modern portfolio theory, the optimal portfolio is determined solely by an individual's risk preferences.[10] 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 proxy risk capacity through demographic data and wealth proxies. 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. We use China, Germany, and the United States to capture

[10] Tobin's two-fund separation theorem, which is still the foundation of portfolio allocation theory and the core of undergraduate Finance courses, states that all investors hold some combination of the risk-free asset and a risky market portfolio, which is defined as the most efficient combination of all available securities (Merton, 1969; Tobin, 1958). The exact proportions of the two components are governed by an individual's risk preferences, with more risk-averse investors holding higher proportions of the risk-free asset.

important developed and emerging economies. There are a number of ways an investor's home country may affect the suggested portfolios. On the one hand, investments in domestic securities may be beneficial due to the lack of foreign exchange rate risk and superior access to and interpretability of information (Coval and Moskowitz, 1999). On the other hand, excessively investing in domestic securities creates cluster risk as both capital and labor income are affected by the economic development in the investor's home country.

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 fulfils the client's sustainability preferences [...]" (ESMA, 2018, p.23). In practice, this is usually achieved by investing in funds that explicitly exclude securities of firms whose business model is inconsistent with environmental, social or governance-related sustainability.

We further include some investor profiles with an explicit reference to 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 differences, 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 advice providers are typically required to ensure their clients understand the nature of the risks associated with the investment products they are recommended.

Table B1 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 models 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 request text:

> 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 depicted in Table B1 in the Appendix. Data was collected between January 18, 2024, and January 23, 2024.[11] Portfolio recommendations including security ticker and portfolio weights are compiled for each model-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). 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.[12] While we do not specify a specific analysis period, this potentially introduces a look-ahead bias in the performance measures as LLMs may observe large parts of the historical performance in their training data, which they could use to recommend ex-post optimal portfolios. 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 a model's cutoff from information in two additional analyses. First, we account for a model's information cutoff date when we investigate the relationship between model 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. While the results do not convincingly suggest that LLMs draw on past performance when recommending portfolios, the performance results should still be taken with a grain of salt and should only be interpreted as a rough measure of fit between historical performance and assumed risk.

[11] Each prompt is only requested once. To ensure consistent responses, we restrict our analyses to models for which the temperature could be set to zero, which ensures that a model provides the same answer to the same prompt.

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



Figure 2: Illustration: GPT-3.5 Turbo portfolio suggestion

3.3 Large language models

We use a total of 32 large language models for our analysis. We initially compiled a list of 76 models featured on common industry leaderboards[13] and added 3 recent and highly publicized releases that were not contained on the leaderboards at the time the model list was compiled (Mistral Medium, Gemini, Grok-1). Whenever there are multiple different versions of a particular model, we use the largest (in terms of parameters) and most recent version.[14] 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 models, we were unable to access 41.[15] Finally, we restrict our analysis to models 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 across 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 models, leaving us with 32 models, for which we obtain portfolio recommendations.

Table 2 contains the models we use for our analyses as well as some key characteristics. We characterize models 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 on the value one if a model uses more than 60 billion parameters, which is effectively a median split (15 large models, 17 small models). Defining more granular size categories (e.g., tertile or quartile split) might lead to collinearity in model characteristics. Specifically, the eight largest models are all foundation models published under a proprietary license, which implies that there is no variation in model type among the largest quartile of models (see Table 2).

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

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

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

[14] 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.

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

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

Table B2 in the Appendix reports pairwise correlations between the model characteristics. The correlations imply that foundation models are significantly less likely published under an open-source license and more likely of the largest category (> 75B parameters). This pattern is potentially a reflection of the maintenance of competitive advantages by developers of leading models (OpenAI, Google). Consistent with this notion, larger models are significantly less likely published open-source than smaller models.

3.4 Construction of suitability measures

3.4.1 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 models 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.[16] Finally, if the portfolio recommendations are incomplete or otherwise erroneous, they cannot be implemented in a straight-forward manner. In evaluating the latter, we identify 5 types of potential response errors (sorted from most likely to least likely, see Table 3 for details):

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",

[16] Data is defined as available if we are able to download historical price data from *YahooFinance* using the ticker provided by the LLM. For the data availability measure, we use *YahooFinance* as retail investors likely do not have access to expensive data services such as Bloomberg. For the six-factor models we use in the exposure and performance analyses, we use Bloomberg price data.

Table 2: Model overview

Model	Developer	Model type	Base 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	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 models, size and information cut-offs are not always reported by the developer. In these cases, we triangulate model size from various estimates.

"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 model 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.

3.4.2 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 desribes in detail how we obtain time series from the recommendations generated by the LLMs, which (ideally) contain specific tickers and portfolio weights. 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 were provided by the LLM (even though specifically requested), we assume equal weighting of all suggested products. If ranges were 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 2000 to December 2023 from Bloomberg. We use the adjusted close price series quoted in US dollars, which accounts for stock splits and dividends. Portfolio values are constructed using the suggested portfolio weights and assuming monthly rebalancing.[17]

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:

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

For each portfolio recommended by model 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

[17] Missing product or portfolio returns are replaced with zero-returns. As a robustness test, we additionally conduct the factor model analyses assuming annual rebalancing.

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.[18] For each portfolio, we obtain a measure of risk-adjusted returns (α_{mp}) and a measure of market risk (β_{mp}^{mkt}). For our subsequent analyses, we replace all coefficients with zeros if they are not statistically significant at the 10% level. 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 $\sigma(\epsilon_{mp})$.

3.4.3 Exposure

To assess the fit between portfolio composition and the investor's 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.[19] 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 markets & emerging markets).[20] 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 proxy for total portfolio risk using the volatility of monthly portfolio returns. We proxy for market-wide and idiosyncratic risk using the market beta and idiosyncratic volatility derived from the six-factor model (equation 1).

3.4.4 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

[18] 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

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

[20] 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.

1. Fourth, to account for administrative and management fees, we aggregate product-level total expense ratios (TER) to the portfolio level.

4 Results

Table 3 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 in section 2 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 model m for investor profile p. X_m is a vector containing model characteristics (model 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 descriptive data across models as well as the γ coefficients, while tests of hypotheses 2 and 3 (relationship of specific model features with ability to generate financial advice) will generally be based on the vector of β coefficients.

Table B3 in the Appendix reports pairwise correlation coefficients for the main dependent variables used in our analyses. Some observations are worth noting. First, the implementability measures seem related. Data availability is negatively associated with the incidence of response errors. Second, a higher exposure to developed markets and US securities is associated with higher data availability and fewer response errors. There are two competing explanations for this pattern. It could suggest that LLMs more frequently misreport tickers for emerging market securities, which is why data availability is lower. It could also 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 raw and risk-adjusted returns, but at higher operating costs. Portfolios with more response errors display lower risk-adjusted returns (Sharpe ratios) and higher operating costs. Thus, we will account for heterogeneity in implementability when we investigate performance.

4.1 Implementability

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

Panel A of Table 3 reports summary statistics for the implementability measures. The number of recommended assets is generally moderately low (mean = 4.8, median = 5.0),

Table 3: Summary statistics

	N	Mean	Median	SD	Min	Max
Panel A: Implementability						
No. assets	2,048	4.81	5.00	3.64	1	73
Data available	1,561	0.93	1.00	0.20	0	1
Some response error	2,048	0.47	0.00	0.50	0	1
By type						
Weights do not add up to 100%	2,048	0.34	0.00	0.47	0	1
No recommendation	2,048	0.17	0.00	0.37	0	1
No portfolio weights	2,048	0.07	0.00	0.25	0	1
Wrong ticker	2,048	0.05	0.00	0.16	0	1
No ticker provided	2,048	0.04	0.00	0.18	0	1
Panel B: Exposure						
Equity	1,564	0.67	0.72	0.27	0	1
Fixed income	1,564	0.30	0.24	0.25	0	1
Alternative assets	1,564	0.02	0.00	0.06	0	1
Cash	1,564	0.01	0.00	0.07	0	1
Individual stocks	2,048	0.02	0.00	0.10	0	1
US securities	1,561	0.65	0.70	0.27	0	1
German securities	1,561	0.04	0.01	0.10	0	1
Chinese securities	1,561	0.07	0.01	0.19	0	1
Developed markets	1,561	0.88	0.94	0.20	0	1
Emerging markets	1,561	0.12	0.06	0.20	0	1
Domestic securities	1,264	0.49	0.59	0.37	0	1
Monthly volatility (%)	1,559	2.88	2.78	3.14	0	114
FF6 market beta	1,559	0.51	0.51	0.24	0	1
Idiosyncratic volatility (%)	1,559	0.71	0.79	0.23	-0	1
Panel C: Performance						
Mean monthly return (%)	1,559	0.19	0.15	0.33	-0	8
Annual Sharpe ratio	1,559	0.16	0.20	0.28	-3	1
FF6 alpha (%)	1,559	-0.11	-0.16	0.26	-1	4
Total expense ratio (%)	1,534	0.18	0.12	0.21	0	3

Note: Summary statistics are displayed on the model-profile level (32 models × 64 profiles). Assets depicts the number of assets per model-profile combination. 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 7 issues occurs in a model-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 price data from Bloomberg. Total expense ratios are computed as weighted averages of product-level ratios.

but is at times as high as 73. Data is available for a majority of tickers (93% 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% (32%).[21] The second most frequent error was the lack of a specific recommendation (17%), which occurred either because the models generated a generic response (mostly suggesting a low-cast passive strategy) or because no response was received for a prompt through the API. Decomposing this error type by model, we find that two models failed to generate specific recommendations for all 64 profiles, while 13 models always provided specific recommendations and the remaining models 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, we show in an additional analysis that explicitly addressing the potential errors in the prompt significantly reduces incidence errors. Specifically, we use simple prompts in our main analyses to obtain results that are comparable across models. Related studies have documented the potential performance improvement achieved by clever 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 this end, we first collect data for profiles 1 through 56 and for each available model using the exact same prompts as in the original data collection (*original* prompts).[22] We then use five different variations on the original prompt, which explicitly reference each of the five error types we observe (*correction* prompts).[23] We then regress the incidences of the various error types on a dummy variable indicating whether a correction prompt has been used while controlling for model and profile characteristics. The resulting coefficients are reported in Tables B16 (OLS) and B17 (logistical regression) in the Appendix. 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. Thus, we conclude that most LLMs are generally able to generate financial advice that is in fact implementable.

[21] 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%.

[22] Of the 32 original models, 19 models were still available at the time of the second data collection. For each of the 1,064 recommendations (19 models × 56 profiles), we obtain data for 6 different prompts (1 original and 5 correction prompts). Data were collected from August 22 to 27, 2024.

[23] For instance, the correction prompt for error type 1 (no portfolio weight specified) consists of the original prompt and the sentence "Please make sure to specify the portfolio weights for each recommended financial product.".

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

To assess how different model and profile characteristics correlate with the implementability of the LLMs' financial advice, we perform multivariate regressions of the implementability measures on model 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 4 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 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 small increase in the number of assets (0.5 assets) from a moderate baseline (4.3) should not affect implementability materially. The difference in error incidence 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 B5 in the Appendix). We find no significant difference in data availability conditional on correctly specified tickers.

To avoid our results being driven by outlier models, we re-estimate the regressions with a restricted sample excluding models with either 100% response errors or 0% data availability (3 models, see Table B5 in the Appendix). The results remain qualitatively unchanged. 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 models of different sizes in the three implementability dimensions. The results suggest that larger models generate portfolio suggestions that are more easily implemented than smaller models. 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. The likelihood of response errors is negatively related to model size. Specifically, the incidence of response errors is 13 percentage points lower for larger models. The corresponding logit specification suggests a decrease of 44% in the error odds-ratio. The results are replicated when excluding models with either 100% response errors or 0% data availability (see Table B5 in the Appendix).

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

Finally, the coefficients on the investor profile characteristics are mostly in line with expectations. Specifically, more risk-tolerant profiles are recommended more assets, potentially related to the higher variety in equity funds vis-à-vis bond funds (portfolios suggested to risk-

tolerant profiles tilt more towards equity, see Table 5). Besides, suggestions for German and Chinese investor profiles display lower data availability and a higher likelihood of response error, which is likely due to the higher prevalence of US securities in global capital markets and the training data. 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, which do not feature as prominently in the training data as conventional investment products. The results are robust to the exclusion of outlier models (see Table B5 in the Appendix).

To test whether the higher error incidence among German, Chinese, and sustainability-oriented investor profiles is indeed driven by a lack of available investment products in the training data, we run an experiment using ChatGPT 4o, the most recent version of OpenAI's chatbot, which we do not use in our main analysis. Specifically, we prompt ChatGPT 4o for the number of available securities by domicile country (China, Germany, United States) and instrument type (equity ETFs, sustainable equity ETFs, fixed-income ETFs, and stocks). To account for the fact that many Chinese mainland companies' stocks are listed in Hong Kong, we additionally allow securities to be domiciled in Hong Kong. We also prompt for securities domiciled in the Euro area as a whole to account for Euronext listings. Table B6 in the Appendix reports the resulting figures, which suggest that the number of US-based ETFs exceeds those domiciled in Germany or China for all ETF types and by a wide margin (e.g., 2,600 equity ETFs domiciled in the United States vs. 175 domciled in Germany and 550 domiciled in China). The results change slightly when we look at individual stocks or the entire Euro area. This pattern confirms our suspicion that the higher incidence of response errors among Chinese, German, and sustainability-oriented investors is driven by a lack of available information in the training data.

Table 4: Implementability: Multivariate regressions

	(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	Logit	Logit
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 one of the 8 implementability issues described above. Omitted categories are "fine-tuned" for model type, "≤ 60B" for model 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 ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

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). We further study whether model 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 (i) compare the exposure measures of the LLM-generated portfolios to those of un-advised and advised retail investors investigated in related studies (Bhattacharya et al., 2024, 2012), and (ii) assess whether the relationship between investor characteristics and exposure measures coincides with modern portfolio theory (Lintner, 1965).

Panel B of Table 3 reports summary statistics for the exposure variables.[24] 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. These figures are broadly consistent with figures reported in an audit study involving German retail investors (Bhattacharya et al., 2012, p. 983). On average, the Markowitz-type optimization algorithm 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.

With respect to markets, the average portfolio allocates 65% to US securities, 4% to German securities, and 7% to Chinese securities. Correspondingly, the average share invested in the investor's home country is 49% across all profiles.[25] The average portfolio allocates 88% to developed market securities and 12% to emerging market securities.

To assess whether portfolio exposure is suitable for a specific investor profile, we perform multivariate regressions of various exposure variables on model and profile characteristics (as specified in equation 2). As dependent variables, we use the share of equity and fixed income securities as measures of asset class exposure.[26] We use exposure to developed markets, US, and domestic securities as measures of geographical exposure. Finally, we measure exposure to risk using total risk (monthly volatility), market-wide risk (FF6 beta), and portfolio-specific risk (idiosyncratic volatility). The results of the baseline specification are reported in Table 5.

Across models, high-risk-tolerance investors were recommended portfolios with a 30 percentage point higher equity share ($p < 0.01$) than low-risk-tolerance investors, which translates

[24] Table B9 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.

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

[26] We do not use the exposure to alternative assets, mixed-asset securities or cash due to their marginal average shares (see Table 3).

Table 5: Determinants of asset class, market and portfolio risk exposure

	Asset classes			Markets			Risk		
	(1) Equity share	(2) Fixed income	(3) Ind. stocks	(4) Developed markets	(5) US securities	(6) Domestic securities	(7) Monthly vola. (%)	(8) FF6 β_{mkt}	(9) Idios. vola.
Model characteristics									
Foundation model	-0.018	0.017	0.002	0.008	0.031**	0.033**	-0.159	0.011	-0.003
	(0.012)	(0.012)	(0.005)	(0.009)	(0.014)	(0.015)	(0.200)	(0.011)	(0.012)
Size > 60B	-0.043***	0.051***	0.003	-0.011	-0.022	0.027*	-0.145	0.029***	0.034***
	(0.011)	(0.011)	(0.004)	(0.009)	(0.014)	(0.016)	(0.147)	(0.010)	(0.011)
Open-source	-0.018	0.035***	-0.001	0.047***	0.037**	-0.067***	-0.137	-0.003	0.003
	(0.014)	(0.014)	(0.005)	(0.010)	(0.016)	(0.018)	(0.088)	(0.012)	(0.013)
Investor characteristics									
Age = 45	-0.049***	0.051***	0.004	-0.004	0.000	0.013	-0.137	-0.015	-0.016
	(0.013)	(0.013)	(0.005)	(0.012)	(0.017)	(0.019)	(0.087)	(0.013)	(0.013)
Age = 60	-0.096***	0.082***	0.017***	0.008	0.025*	0.040**	0.100	-0.073***	-0.082***
	(0.013)	(0.012)	(0.005)	(0.009)	(0.014)	(0.016)	(0.206)	(0.011)	(0.012)
Risk tolerance = high	0.300***	-0.286***	0.028***	-0.039***	0.002	-0.004	1.389***	0.257***	0.116***
	(0.011)	(0.010)	(0.004)	(0.008)	(0.013)	(0.014)	(0.154)	(0.010)	(0.011)
Home country = US	-0.006	0.012	0.008*	-0.002	-0.006		0.069	0.003	-0.006
	(0.014)	(0.013)	(0.004)	(0.006)	(0.015)		(0.071)	(0.013)	(0.013)
Home country = Germany	-0.008	0.003	0.019***	-0.028***	-0.158***	-0.604***	0.362	-0.030*	-0.014
	(0.017)	(0.016)	(0.006)	(0.008)	(0.020)	(0.015)	(0.396)	(0.016)	(0.016)
Home country = China	0.025	-0.040**	0.030***	-0.283***	-0.227***	-0.418***	0.372***	-0.074***	-0.172***
	(0.018)	(0.017)	(0.007)	(0.022)	(0.023)	(0.023)	(0.144)	(0.017)	(0.020)
Sustainability = Yes	0.089***	-0.081***	-0.002	0.002	0.008	-0.011	0.338*	0.003	-0.017
	(0.012)	(0.011)	(0.005)	(0.010)	(0.014)	(0.016)	(0.195)	(0.011)	(0.012)
Constant	0.577***	0.373***	-0.010*	0.926***	0.677***	0.712***	2.144***	0.404***	0.714***
	(0.022)	(0.021)	(0.006)	(0.015)	(0.026)	(0.026)	(0.136)	(0.020)	(0.022)
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	0.050	0.332	0.164

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 model type, "$\leq$ 60B" for model 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 ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$).

to significantly higher portfolio risk (using all three risk measures). This difference is generally consistent with Bhattacharya et al. (2012), whose algorithm recommends a 25 percentage points difference in equity share for the most versus least risk tolerant investors.[27] We further find that portfolios recommended to high-risk-tolerance investors tilt more towards emerging markets (4 percentage points 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 toward retirement, we also find that equity exposure is decreasing in age (10 percentage points difference for 60-year-olds vs. 30-year-olds, $p < 0.01$) and fixed-income exposure is increasing in age (8 percentage points difference for 60-year-olds vs. 30-year-olds, $p < 0.01$), which translates to lower market and portfolio-specific 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. Corroborating this notion, there is no statistically significant difference in monthly volatility between the portfolios recommended to 60-year-olds and 30-year-olds (though market and portfolio-specific risk are lower). Interestingly, we also find that the oldest investors are recommended a higher share of individual stocks. In a lot of cases, this is due to the models recommending stocks known to pay high and consistent dividends as retirement approaches. All results are robust to the inclusion of implementability measures as covariates (see Table B7 in the Appendix) and to using region-specific factor portfolios and an annual rebalancing frequency (see Table B10 in the Appendix).

According to modern portfolio theory, an investor's risk tolerance should be the main determinant of portfolio risk (Lintner, 1965). 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 (cf. Blaseg and Hornuf, 2024). The weights reflect the relative contribution of each investor characteristic to the total predicted variance in the exposure variables. For robustness, we further compute 95% confidence intervals for each weight based on 10,000 bootstrap samples. Table 6 displays 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, 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, 87% of the explained variation in monthly volatility, 89% of the explained variation in market risk, and 38% of the explained variation in idiosyncratic volatility. It dominates any other investor characteristic in the first three specifications. Only

[27] 63% recommended to high-risk-tolerance investors, 38% recommended to low-risk-tolerance investors (Bhattacharya et al., 2012, Table 2).

in idiosyncratic volatility regressions, it is dominated by the Chinese investor dummy. This pattern persists when we use region-specific factor portfolios (see Table B11 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. Generally, the results suggest that risk tolerance is the most important determinant of portfolio risk, which is consistent with the prescriptions of modern portfolio theory.

Taken together, the relation between the exposure variables and investor characteristics across models suggest that — generally — LLMs are capable of providing financial advice that takes into account investors' specific circumstances. All correlations are in line with principles of sound financial advice and consistent with the recommendations of a rule-based Markowitz optimizer (Bhattacharya et al., 2012). Importantly, in line with the prescriptions of modern portfolio theory, risk tolerance is the main determinant of the portfolios' equity share and risk.

Result H1.2. *LLMs generally recommend portfolios that take into account the investor's risk tolerance and capacity.*

In addition to the baseline regression, we assess the degree to which different model features affect the sensitivity of our exposure measures to the individual circumstances of the investor profiles. Since 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 between investors with high and low risk tolerance. Positive coefficients suggest that foundation models and larger models are more sensitive to an investor's risk tolerance than fine-tuned models and smaller models.

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 model types of 7 percentage points

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
Model characteristics												
Foundation model	0.00	0.00	[0.00, 0.00]	0.00	0.01	[0.00, 0.00]	0.00	0.00	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]
Size > 60B	0.01	0.01	[0.00, 0.01]	0.00	0.01	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]	0.00	0.03	[0.00, 0.01]
Open-source	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]
Investor characteristics												
Age = 45	0.00	0.01	[0.00, 0.00]	0.00	0.01	[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, 0.00]	0.02	0.05	[0.01, 0.03]	0.02	0.14	[0.01, 0.04]
Risk tolerance = high	0.32	0.84	[0.28, 0.35]	0.05	0.87	[0.02, 0.26]	0.30	0.89	[0.26, 0.34]	0.06	0.38	[0.04, 0.09]
Home country = US	0.00	0.01	[0.00, 0.01]	0.00	0.01	[0.00, 0.00]	0.00	0.01	[0.00, 0.01]	0.01	0.05	[0.00, 0.01]
Home country = Germany	0.00	0.00	[0.00, 0.00]	0.00	0.02	[0.00, 0.00]	0.00	0.00	[0.00, 0.00]	0.00	0.02	[0.00, 0.01]
Home country = China	0.00	0.01	[0.00, 0.01]	0.00	0.02	[0.00, 0.01]	0.01	0.03	[0.00, 0.02]	0.06	0.37	[0.04, 0.09]
Sustainability = yes	0.03	0.07	[0.01, 0.04]	0.00	0.05	[0.00, 0.01]	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			0.06			0.34			0.17	

Note: The table reports standardized dominance statistics for regressions of equity share (column 1), volatility (column 2), six-factor market beta (column 3), and idiosyncractic volatility (column 4) on the various independent variables.

is statistically significant at the 1% level. The higher sensitivity is also reflected in higher market and idiosyncratic risk.[28] Using the difference in equity shares observed in professionally managed portfolios as a benchmark (Fieberg et al., 2023), the greater sensitivity is considered more suitable in our context.[29]

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

Regarding model size, our results suggest that larger models are more sensitive to differences in investors' risk preferences: While the difference in equity shares between high- and low-risk-tolerance investors is only 26 percentage points for small models, it increases to 34 percentage points for large models (increase of 8 percentage points, $p < 0.01$). Thus, while larger models tilt more towards fixed income across 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 market risk ($p < 0.05$), but not idiosyncratic volatility.[30] Taken together, we conclude that larger models recommend portfolios that are more suitable for the investors' individual circumstances.

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

[28] Both results are robust to using region-specific factor models and assuming annual rebalancing (see Table B12 in the Appendix).

[29] In Fieberg et al. (2023), we document a difference in equity shares between investors with high and low risk tolerance of 45 percentage points (90 vs. 55%) in the professionally managed portfolios.

[30] Both results are robust to using region-specific factor models and assuming annual rebalancing (see Table B12 in the Appendix).

Table 7: Impact of risk tolerance on equity share and portfolio risk, by model 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***	1.373***	1.464***	0.220***	0.235***	0.084***	0.111***
	(0.015)	(0.015)	(0.374)	(0.304)	(0.015)	(0.014)	(0.016)	(0.016)
Model type								
Risk tolerance = high × original model	0.071***		0.026		0.062***		0.052**	
	(0.021)		(0.385)		(0.020)		(0.021)	
Model size								
Risk tolerance = high × Size > 60B		0.075***		-0.148		0.043**		0.010
		(0.021)		(0.314)		(0.019)		(0.021)
Constant	0.599***	0.595***	2.152***	2.110***	0.423***	0.414***	0.730***	0.716***
	(0.023)	(0.023)	(0.185)	(0.164)	(0.020)	(0.020)	(0.023)	(0.022)
Obs.	1,564	1,564	1,559	1,559	1,559	1,559	1,559	1,559
Adj. R2	0.378	0.379	0.049	0.049	0.336	0.334	0.167	0.164
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), a dummy variable indicating an open source-model (vs. proprietary), and three dummy variables indicating model size: ≤60B indicates models with more than 7B, but at most 60B parameters. ≤75B indicates models with more than 60B, but at most 75B parameters. ≥75B indicates models with more than 75B parameters. Models with at most 7B parameters are the baseline group. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

4.3 Performance

In this section, we investigate the historical performance of the LLMs' portfolio recommendations. Since the performance evaluations are by necessity based on backtesting, it is conceivable that LLMs base their portfolio allocation decisions on previous returns (look-ahead bias). To account for this, we include a model's information cut-off in our regression analyses. In addition, we run an experiment around an update to ChatGPT to assess whether it adjusts its portfolio recommendations to recent price developments.

Panel C of Table 3 displays summary statistics of the performance measures. The average monthly excess return is 0.19%, the average annual Sharpe Ratio is 0.2, six-factor alphas are slightly negative on average (-0.11%) and range from -1% to 4%.[31] Compared to the overall market, which experienced average monthly returns of 0.46% and a Sharpe ratio of 0.46 over the same horizon, the average portfolio underperforms. TER is 0.18% on average, which suggests that most portfolios employ low-cost passive investment vehicles. The average performance figures are qualitatively unchanged when we employ region-specific factor portfolios or assume annual rebalancing (see Table B9 in the Appendix).

[31] 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.

We next compare our performance estimates to those reported in related studies. Broadly in line with our estimates, Dahlquist et al. (2017) find that the average (Swedish) retail investor generates a return of 4.03% per year (which corresponds to roughly 0.33% per month) and risk-adjusted returns (using a hybrid 4-factor model) are not distinguishable from zero. Professional financial advisors do not necessarily contribute to improving retail investors' performance: Hoechle et al. (2017), who analyze the effect of financial advice on a trade-by-trade basis, find that advised trades generate negative monthly abnormal returns to a five-factor model (both buy and sell transactions).[32] In a previous version of this study, we find that a professionally managed digital advice offering generates significantly negative six-factor alphas (approximately -20 bps, cf. Fieberg et al., 2023). Taken together, the performance measures generally support the notion that LLMs have the capacity to recommend portfolios whose performance is in line with their risk exposure. However, since the performance figures may be affected by look-ahead bias, they may overestimate out-of-sample performance measures.

Result H1.3. *LLMs recommend portfolios whose historical performance is in line with their risk exposure.*

To assess how different model and profile characteristics correlate with the performance measures, we perform multivariate regressions of various exposure variables on model and profile characteristics (according to equation 2). To address concerns over look-ahead bias, we additionally account for the recency of a model's training data. Most LLMs are trained using internet data up to some information cut-off date. For instance, OpenAI's GPT-4-Turbo is trained on information dated December 2023. On the other hand, some models (e.g. Perplexity AI's PPLX-70B model) are provided with regular information updates that are intended to maintain the timeliness of the model's responses. To account for this, we include a dummy variable indicating that a model's training data cutoff falls within 6 months from the time of our analysis (April 2024).[33]

Table 8 reports the results of the main regression specifications. Regarding hypothesis 2, we find that foundation models generate portfolios with significantly lower Sharpe ratios ($p < 0.01$), 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 misspecified ticker) are systematically different in terms of performance. Table B8 in the Appendix reports the resulting coefficients. The results are in line with the main results. Second, we use region-specific factor portfolios and allow for annual rebalancing,

[32] More specifically, they find that risk-adjusted performance was significantly worse (better) for advised versus independent purchase (sell) decisions.

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

which does not affect our main results (see Table B13 in the Appendix). Taken together, 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 lower.

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 models recommend portfolios with slightly lower excess returns ($p < 0.1$ when information cutoff is considered) and Sharpe ratios ($p < 0.01$), but no difference in FF6 alphas. When taking into account data availability and response errors (Table B8), only the difference in Sharpe ratios persists ($p < 0.05$). The results are unchanged when we use region-specific factor portfolios (see Table B13 in the Appendix). Taken together, we find no evidence of superior performance of larger models. If anything, risk-adjusted performance as measured by annual Sharpe ratios is lower.

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

The results reported in Table 8 suggest that LLMs with a more recent information cutoff generate lower Sharpe ratios ($p < 0.01$) and higher operating costs ($p < 0.01$) than LLMs with no exposure to recent information. Both results are robust to including the data availability and response error variables from the implementability analyses (see Table B8 in the Appendix). When implementability is taken into account, six-factor alphas are slightly higher for models with access to more recent information (3 bps, $p < 0.1$). The difference increases to 4 basis points ($p < 0.1$) when using region-specific factor portfolios and 5 basis points ($p < 0.01$) when using region-specific portfolios and assuming annual rebalancing (see Table B13 in the Appendix). These results are somewhat surprising as we would have expected 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 to be driven by news coverage of particularly well-performing stocks. Consistent with this notion, the most frequently recommended individual stocks include — aside from well-known American tech stocks — stocks with a good recent performance such as Netflix and Nvidia.[34]

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 (Fieberg et al., 2023). At that time, the model's knowledge cutoff was September

[34] 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).

Table 8: Determinants of portfolio performance and fees

	(1) Mean monthly return (%)	(2) Mean monthly return (%)	(3) Sharpe ratio	(4) Sharpe ratio	(5) FF6 α	(6) FF6 α	(7) Total exp. ratio (%)	(8) Total exp. ratio (%)
Model characteristics								
Foundation model	-0.020	-0.027	-0.028*	-0.050***	0.017	0.028	-0.007	0.018
	(0.021)	(0.024)	(0.015)	(0.019)	(0.017)	(0.020)	(0.011)	(0.015)
Size > 60B	-0.018	-0.030*	-0.017	-0.040***	-0.014	-0.007	-0.021**	0.008
	(0.015)	(0.018)	(0.012)	(0.013)	(0.012)	(0.014)	(0.010)	(0.010)
Open-source	-0.032*	-0.006	-0.022	0.008	-0.044**	-0.012	-0.024**	0.009
	(0.016)	(0.021)	(0.016)	(0.019)	(0.017)	(0.021)	(0.011)	(0.012)
Cutoff within 6 months		-0.004		-0.057***		0.023		0.046***
		(0.014)		(0.018)		(0.015)		(0.014)
Investor characteristics								
Age = 45	-0.019	-0.008	-0.024	-0.011	0.006	0.017	0.009	-0.004
	(0.014)	(0.016)	(0.015)	(0.017)	(0.014)	(0.015)	(0.013)	(0.014)
Age = 60	0.013	0.025	-0.049***	-0.042**	0.047***	0.058***	0.014	0.016
	(0.021)	(0.026)	(0.016)	(0.018)	(0.016)	(0.019)	(0.012)	(0.014)
Risk tolerance = high	0.212***	0.219***	0.199***	0.203***	0.095***	0.096***	0.003	-0.000
	(0.016)	(0.019)	(0.013)	(0.015)	(0.013)	(0.015)	(0.010)	(0.012)
Home country = US	0.021*	0.016	0.006	0.002	0.027**	0.022	-0.001	0.003
	(0.013)	(0.015)	(0.015)	(0.018)	(0.012)	(0.015)	(0.010)	(0.011)
Home country = Germany	0.011	0.001	-0.046**	-0.070***	-0.002	-0.015	0.084***	0.068***
	(0.032)	(0.041)	(0.018)	(0.021)	(0.017)	(0.019)	(0.015)	(0.013)
Home country = China	-0.029	-0.042	-0.110***	-0.111***	0.070***	0.053*	0.141***	0.130***
	(0.025)	(0.030)	(0.021)	(0.021)	(0.025)	(0.031)	(0.018)	(0.021)
Sustainability = Yes	-0.003	-0.003	0.007	0.007	-0.021	-0.034**	0.058***	0.063***
	(0.019)	(0.023)	(0.013)	(0.014)	(0.014)	(0.016)	(0.010)	(0.011)
Constant	0.123***	0.106***	0.147***	0.177***	-0.168***	-0.217***	0.142***	0.057***
	(0.023)	(0.034)	(0.022)	(0.031)	(0.023)	(0.036)	(0.016)	(0.021)
Obs.	1,559	1,186	1,559	1,186	1,559	1,186	1,534	1,166
Adj. R2	0.102	0.094	0.153	0.170	0.053	0.050	0.100	0.092

Note: Omitted categories are "fine-tuned" for model type, "$\leq$ 60B" for model 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 ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$).

2021. After OpenAI released an updated version of GPT-4 drawing on data up until January 2022, we collect 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 knowledge 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, but not the initial recommendations 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 reduce 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 in a similar manner as a human advisor would. 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 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, 2024a).

We focus on two distinct biases that may occur in the context of LLM-generated financial advice. First, we find in an earlier version of this paper that portfolio recommendations generated by ChatGPT suffer from home bias, i.e., an excessive allocation to domestic securities (Coval and Moskowitz, 1999; Fieberg et al., 2023). While strategies emphasizing investments in local securities may benefit from privileged information access, they generally result in under-diversified portfolios, especially if the correlation between labor and capital income is taken into account. Second, we study whether LLMs inherit gender-based discrimination. A related study shows that social media algorithms inadvertently target male users when placing 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, 2024b). In the financial advice context, studies document lower stock market participation and risk tolerance among women than men (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 risk tolerance and capacity 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 Morgan Stanley Capital International (MSCI) All-Country World Investable Markets Index as proxies for m_i.[35] 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 German and Chinese securities make up 2% each.[36] Comparing the shares of domestic securities in the portfolios of Chinese, German, and US investors reveals a substantial home bias across home countries. 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 9). German securities account for 11% of all assets and 9% of the equity portion of German investors' portfolios, suggesting a weight-gap of 7 percentage points (see Panel A of Table 9). US securities account for 72% of all assets and 73% of the equity portion of US investors' portfolios, suggesting a weight-gap of 11 percentage points (see Panel A of Table 9). 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 (by including only recommendations where all tickers were correctly specified, panel B) and across risk tolerance levels (panel C).

To compare the severity of the home bias across (home) 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 normative weight. Using this specification, the home bias is most severe for China, with the actual weight exceeding the baseline weight by between 7 and 9 times. 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

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

[36] 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, US securities feature less prominently (49%), whereas Chinese securities feature more prominently (14%) and the German share remains unchanged (2%). Our results are qualitatively unchanged when using this alternative benchmark specification.

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.[37]

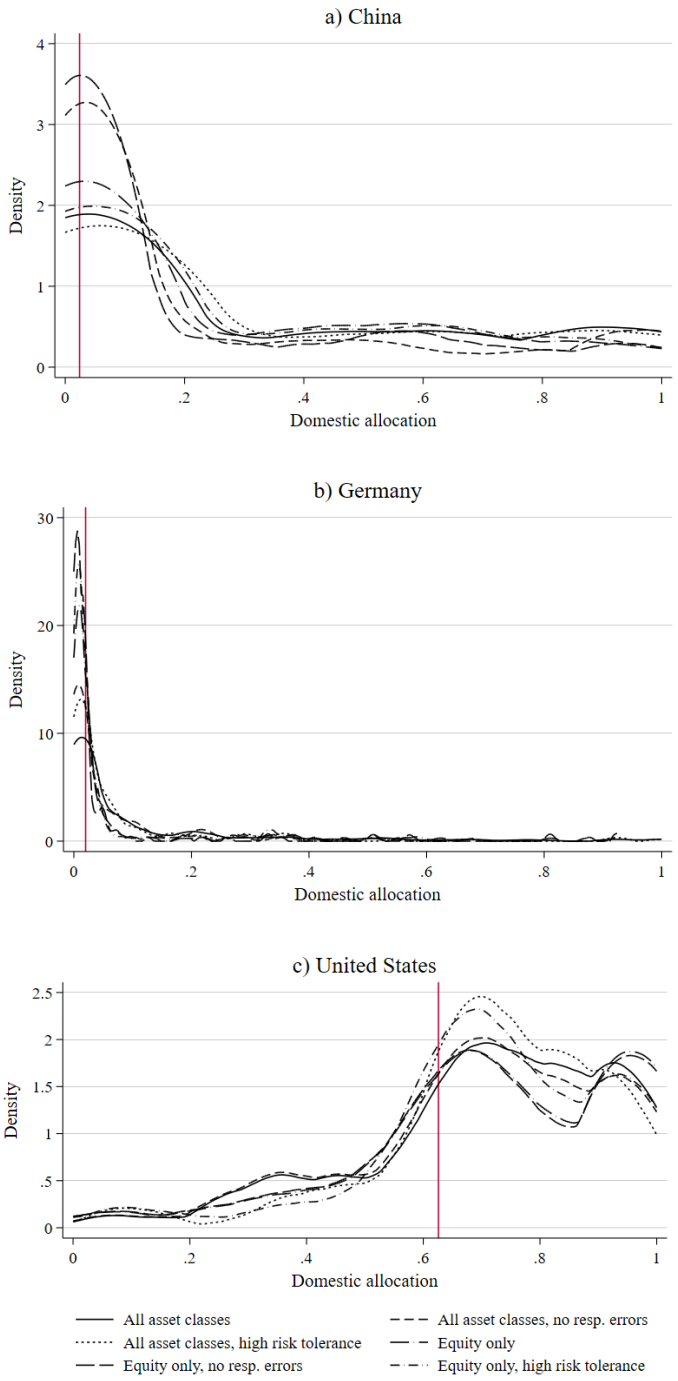
To gauge the distribution of excessive domestic allocation, Figure 3 displays density plots for the domestic allocation (for various sub-samples) by home country. The vertical red lines depict the baseline weights m_i. The density plots confirm our findings from the various home bias measures that while there is a home bias for all home countries, it is highly pronounced for the Chinese and US profiles and less obvious for German profiles, for which the mode of the distribution is below the benchmark weight. This might be a consequence of the use of European rather than national stock and bond funds for domestic exposure in the case of German investors.

One potential explanation for the home bias is that capital cannot easily be moved across international borders (Bekaert and Wang, 2009). Among the three countries used in our analyses, Germany and the United States 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 openness index, cf. IMF, 2016). Thus, if the portfolios recommended to Chinese investors include international securities 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. Since the average international exposure among Chinese investors is 71% (median 93%) and the average US exposure is 50% (median 58%), the true home bias for Chinese investors is likely more pronounced.

While the home bias documented in the LLMs' recommendations is substantial, it is no worse than the home bias documented in 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). 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).

[37] Table B14 in the Appendix displays coefficients from regressions of the home bias measures on model and profile characteristics. The results suggest that larger models' recommendations suffer more from home bias than smaller models'.

Figure 3: 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 the sub-samples listed in Table 9.

Table 9: 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	0.72	703	0.22	0.73	682	0.24	0.63	0.11	11.56***	0.17	0.28
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	0.71	621	0.22	0.73	606	0.24	0.63	0.10	10.47***	0.16	0.27
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	0.72	356	0.20	0.72	354	0.21	0.63	0.09	8.31***	0.15	0.25

Note: The table reports the portfolio weights assigned to domestic securities by home country of the investor profile. The first (second) column reports the average domestic allocation d_i, the number of portfolios, and the standard deviation of the domestic allocation across 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 asterisks ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$). 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.

4.4.2 Gender-based discrimination

Eight of the investor profiles we define explicitly vary the investor's gender (see Table B1 in the Appendix). Besides gender, the profiles vary with respect to risk tolerance (high vs. low) and age (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 have been linked to gender in previous literature (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; d'Astous et al., 2024) and investigate gender differences in the domestic portfolio weights and shares of individual stocks.

Table 10 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 averages. 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 across 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 B15

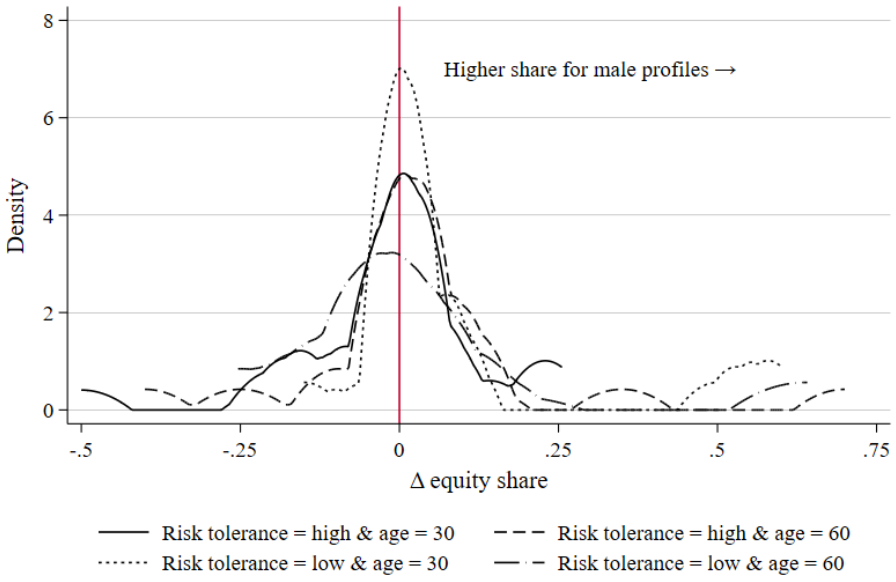
in the Appendix shows, we find no model-specific differences in exposure to markets or risk components in recommendations to male and female investors.

Table 10: Portfolio exposures by gender

		(1) Male		(2) Female		(1) - (2)	
		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	0.69	24	0.66	26	0.03	0.78
Risk tolerance = high	Age = 60	0.72	24	0.70	24	0.03	0.15
Risk tolerance = low	Age = 30	0.43	25	0.42	22	0.01	0.44
Risk tolerance = low	Age = 60	0.32	23	0.28	27	0.04	0.85
Panel C: Idiosyncratic volatility							
Risk tolerance = high	Age = 30	0.84	24	0.80	26	0.03	0.17
Risk tolerance = high	Age = 60	0.84	24	0.81	24	0.03	0.01
Risk tolerance = low	Age = 30	0.73	25	0.76	22	-0.03	-0.10
Risk tolerance = low	Age = 60	0.63	23	0.57	27	0.06	0.48
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 from a non-parametric rank test for differences in means ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

Figure 4: Distribution of gender difference in equity share



Note: The figures display density plots of the difference in equity shares in portfolios recommended to males versus females. 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.

5 Conclusion and outlook

In this study, we explore the potential of LLMs to provide financial advice. We elicit 2,048 portfolio recommendations for 64 investor profiles from 32 different models. To assess the suitability of these portfolios, we investigate their implementability, the appropriateness of the portfolios' exposures given investor characteristics, 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 models are better suited to generate portfolio recommendations than fine-tuned and smaller models.

In general, our results suggest that LLMs may be used for financial advice in the future as most models recommended portfolios that (i) can in fact be implemented, (ii) take into account the individual circumstances of the investor, and (iii) whose historical performance is in line with its exposure to commonly priced risk factors. Turning to heterogeneity with respect to model features, our results suggest that foundation models and larger models generate financial advice which is more easily implemented and more suited to the investor's circumstances than fine-tuned models and smaller models. We do not observe any systematic performance differences between the various model features.

Given these encouraging results, what are likely future applications of LLMs in the financial advice context? First, there are some obvious risks associated with the unrestricted use of LLMs in financial advice provision as 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; Lo and Ross, 2024). 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 digital advice systems assign pre-determined weights to a set of pre-selected investment products (Rühr et al., 2019). Narrowing the range of potential portfolio recommendations could be achieved through 'knowledge injection' (Ovadia et al., 2023). One way of operationalizing this is to adjust the models' weights to the financial advice context through fine-tuning. Fine-tuning models to the financial context is challenging as high-quality financial data is hard to obtain (Lo and Ross, 2024). Besides, evidence on the performance of Finance-specific models is sparse and conflicting. On the one hand, BloombergGPT, a model trained in part on Bloomberg's proprietary data, performs better at Finance-specific sentiment analysis than comparable models (Wu et al., 2023). On the other hand, models fine-tuned to the financial context (BloombergGPT, FinBERT) have been shown to underperform cutting-edge general-purpose models such as GPT4 in analyzing financial text (Li et al., 2023). Another way of injecting information specific to the financial advice context is in-context learning, mostly operationalized through retrieval-augmented generation (RAG). By (implicitly) adding context to each query, a model'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 model will then prioritize as input for the provision of financial advice.

A second challenge associated with the use of LLMs regards ethical advice giving. A crucial issue is LLMs' tendency for sociopathic behavior: We described the capacity to simulate a personal relationship as a potential benefit of employing LLMs for financial advice. This may also pose a challenge (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 instance, 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 (in analogy to the Hippocratic oath in medicine) can reduce self-serving advice giving (Weitzel and Kirchler, 2023). Future work will have 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. First and most notably, our performance analyses are based on historical data from 2000 to 2023. We employ this approach because we are interested in the long-term performance of the suggested portfolios. However, since LLMs are trained on large corpuses of texts mostly drawn from the internet, it is conceivable that portfolio recommendations generated by these models take into account the previous price evolution of specific investment products. If

this were the case, our performance measures would overestimate the true potential of the recommendations as they suffer from hindsight bias. While a true out-of-sample test can only be conducted 30 years from now, two of our findings may help alleviate this concern. First, we find that recommendations generated by models with more recent access to the internet do not outperform recommendations of models with less recent access to the internet when measured by simple risk-adjusted performance measures. Second, an experiment involving an update to GPT-4's information base in 2022 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). These findings help ameliorate concerns over potential look-ahead bias in our performance analyses.

References

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.

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.

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 (forthcoming)*.

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.

Chan, W. S. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260.

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.

Dahlquist, M., Martinez, J. V., and Söderlind, P. (2017). Individual investor activity and performance. *Review of Financial Studies*, 30(3):866–899.

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.

Dowling, M. and Lucey, B. (2023). Chatgpt for (finance) research: The bananarama conjecture. *Finance Research Letters*, 53:103662.

Drake, M. S., Roulstone, D. T., and Thornock, J. R. (2012). Investor information demand: Evidence from google searches around earnings announcements. *Journal of Accounting research*, 50(4):1001–1040.

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

Feng, Z., Hu, G., and Li, B. (2023). Unleashing the power of chatgpt in finance research: Opportunities and challenges. *Available at SSRN 4424979*.

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

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

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.

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.

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.

Hoechle, D., Ruenzi, S., Schaub, N., and Schmid, M. (2017). The impact of financial advice on trade performance and behavioral biases. *Review of Finance*, 21(2):871–910.

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 2008].

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.

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 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. *Journal of Finance*, 7(1):77–91.

Merton, R. C. (1969). Lifetime portfolio selection under uncertainty: The continuous-time case. *The Review of Economics and Statistics*, 51(3):247–257.

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.

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. (2023a). GPT as a financial advisor. *Available at SSRN 4384861*.

Niszczota, P. and Abbas, S. (2023b). 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.

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 2008].

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*.

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

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

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.

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

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

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

The New York Times (2022). The brilliance and weirdness of ChatGPT. https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html. [online; accessed 19 August 2008].

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

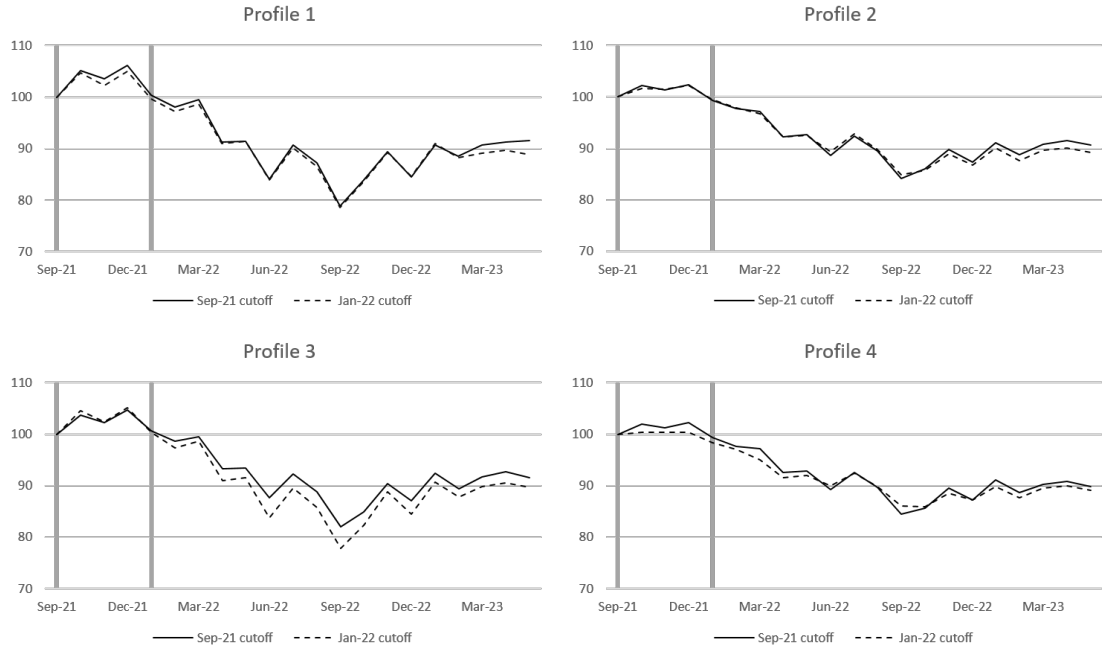
Weitzel, U. and Kirchler, M. (2023). The banker's oath and financial advice. *Journal of Banking & Finance*, 148:106750.

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*.

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*.

Online Appendix A: 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 on 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 cutoffs of the initial and updated recommendations (Sep 21 and Jan 22).

Online Appendix B: Tables

Table B1: 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 B2: Pairwise correlations: Model characteristics

	(1)	(2)	(3)	(4)	(5)
(1) Foundation model	1.000				
(2) Open-source	-0.323*	1.000			
(3) Size $\leq$ 60B	0.131	0.272	1.000		
(4) Size $\leq$ 75B	-0.218	0.234	-0.361**	1.000	
(4) Size $>$ 75B	0.424**	-0.713***	-0.301*	-0.277	1.000
(5) Cutoff within 6 months	-0.220	0.164	-0.200	0.031	-0.236

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

Table B3: Pairwise correlations: 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	0.0299	0.0190	0.00148	0.277***	-0.278***	-0.0596*	-0.0398	-0.0240	1						
(10) FF6 market beta	0.0444	0.145***	-0.164***	0.706***	-0.666***	0.114***	0.103***	-0.0304	0.239***	1					
(11) Idiosyncratic volatility	0.000388	0.186***	-0.163***	0.421***	-0.351***	0.388***	0.184***	-0.134***	0.0398	0.710***	1				
(12) Mean monthly return	0.163***	0.149***	-0.0265	0.343***	-0.355***	0.119***	0.203***	-0.00774	0.815***	0.388***	0.159***	1			
(13) Sharpe ratio	0.198***	0.238***	-0.0931***	0.518***	-0.468***	0.231***	0.289***	-0.00181	0.235***	0.616***	0.492***	0.576***	1		
(14) FF6 alpha	0.209***	0.0620*	0.0330	0.134***	-0.173***	-0.0390	0.216***	0.112***	0.260***	0.121***	-0.144***	0.653***	0.441***	1	
(15) Total expense ratio	0.124***	-0.289***	0.217***	0.0581*	-0.148***	-0.438***	-0.242***	0.00404	0.0211	-0.180***	-0.324***	-0.125***	-0.381***	-0.0137	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 B4: LLM leaguetable

Rank	Model	Implementability		Exposure		Performance		
		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	Jurassic-2 Ultra	0.99	0.23	0.15	-0.02	0.43	0.18	0.16
3	Deepseek	0.99	0.09	0.07	0.48	0.16	-0.13	0.11
4	Solar	1.00	0.31	-0.06	0.31	0.28	-0.01	0.12
5	PPLX-70B	0.98	0.41	0.13	0.28	0.21	0.01	0.13
6	Yi-34B-Chat	0.98	0.20	0.06	0.39	0.19	-0.15	0.08
7	Gemini-Pro	0.89	0.19	0.22	0.39	0.15	-0.09	0.16
8	alpaca-7b	1.00	0.58	-0.03		0.22	-0.09	0.07
9	chatglm3-6b	0.93	0.92	0.21	0.15	0.26	0.04	0.14
10	GPT-3.5-Turbo	0.91	0.31	0.08	0.27	0.20	-0.16	0.13
11	GPT-4-Turbo	0.87	0.14	0.13	0.44	0.12	-0.14	0.16
12	OpenHermes-2p5-Mistral-7B	0.94	0.41	0.05	0.10	0.24	-0.15	0.14
13	LLaMA-20-70B-SteerLM-Chat	0.91	0.45	0.22	0.34	0.18	-0.04	0.23
14	Tulu-2-DPO	0.93	0.13	0.12	0.40	0.10	-0.18	0.19
15	mixtral-8x7b-instruct-v0.	0.92	0.19	0.11	0.32	0.14	-0.17	0.18
16	Mistral-Medium	0.84	0.22	0.22	0.28	0.15	-0.15	0.24
17	Claude 2.1	0.83	0.28	0.16	0.24	0.16	-0.08	0.24
18	falcon-40b-instruct	1.00	0.98	0.10	0.31	0.23	0.01	0.27
19	Platypus2-70B-instruct	0.98	0.59	0.02	0.22	0.19	-0.13	0.14
20	openchat-3.5-1210	0.98	0.41	0.01	0.23	0.17	-0.20	0.16
21	Claude 2.0	0.87	0.27	0.17	0.29	0.05	-0.18	0.21
22	starling-lm-7b-alpha	0.99	0.78	0.21	0.25	0.14	-0.14	0.23
23	Command-Nightly	0.93	0.61	0.23	0.34	-0.02	-0.04	0.36
24	llama-2-70b-chat	0.98	0.63	-0.03	0.34	0.16	-0.17	0.14
25	orca-2-13b	0.99	0.75	0.00	-0.03	0.15	-0.24	0.13
26	Qwen-72B-Chat	0.88	0.44	0.13	0.42	0.08	-0.18	0.15
27	zephyr-7b-beta	0.95	0.38	0.09	0.21	0.09	-0.20	0.24
28	WizardLM	0.89	0.55	0.17	0.43	0.11	-0.19	0.19
29	RedPajama-INCITE-7B-Chat	0.94	0.95	0.10	0.17	0.17	-0.13	0.30
30	Bard-Jan-24-Gemini-Pro	0.79	0.63	0.40	0.53	-0.02	-0.13	0.24
31	StripedHyena-Hessian-7B		1.00					
32	vicuna-33b-v1.3		1.00					
	Average	**0.66**	**0.47**	**0.12**	**0.30**	**0.10**	**-0.13**	**0.18**

Note: The table reports average suitability measures by model. 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.

Table B5: Implementability: Multivariate regressions (robustness tests)

	(1) No. assets	(2) Data available	(3) Some resp. error	(4) Data available	(5) Some resp. error
Model characteristics					
Foundation model	0.510***	-0.018*	-0.109***	-0.717	-0.471***
	(0.166)	(0.009)	(0.025)	(0.514)	(0.104)
Size > 60B	0.117	-0.015	-0.116***	0.220	-0.508***
	(0.152)	(0.011)	(0.024)	(0.395)	(0.105)
Open-source	0.163	0.051***	0.078***	1.594***	0.351***
	(0.152)	(0.013)	(0.027)	(0.454)	(0.118)
Investor characteristics					
Age = 45	-0.183	-0.009	-0.006	-0.478	-0.025
	(0.195)	(0.013)	(0.029)	(0.465)	(0.129)
Age = 60	-0.154	-0.015	0.020	-0.747*	0.090
	(0.208)	(0.010)	(0.025)	(0.451)	(0.113)
Risk tolerance = high	0.918***	0.020**	-0.017	0.699*	-0.076
	(0.169)	(0.009)	(0.022)	(0.363)	(0.098)
Home country = US	-0.001	0.010*	-0.031		-0.142
	(0.201)	(0.006)	(0.032)		(0.140)
Home country = Germany	0.112	-0.217***	0.121***	-17.132	0.518***
	(0.242)	(0.019)	(0.037)	(2724.801)	(0.158)
Home country = China	-0.083	-0.136***	0.098***	-16.929	0.420***
	(0.212)	(0.018)	(0.038)	(2724.801)	(0.158)
Sustainability = Yes	0.158	0.007	0.105***	0.754**	0.454***
	(0.158)	(0.011)	(0.024)	(0.367)	(0.105)
Constant	4.165***	0.966***	0.429***	19.032	-0.314
	(0.317)	(0.017)	(0.046)	(2724.801)	(0.198)
Obs.	1,856	1,500	1,856	824	1,856
Adj. R2	0.016	0.221	0.075		
Pseude R2				0.235	0.061
Model	OLS	OLS	OLS	Logit	Logit

Note: Columns 1 and 2 are estimated using the full portfolio sample. Columns 3, 4, and 5 use the full sample excluding models 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 one of the 8 implementability issues described above. Omitted categories are "fine-tuned" for model type, "$\leq$ 60B" for model 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 B6: Number of securities recognised 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 B7: Determinants of asset class, market and portfolio risk exposure (robustness tests)

	Asset classes		**Markets**			**Risk**		
	(1) Equity share	(2) Fixed income	(3) Developed markets	(4) US securities	(5) Domestic securities	(6) Monthly volatility (%)	(7) FF6 β_mkt	(8) Idiosyncratic volatility
Model characteristics								
Foundation model	-0.020	0.018	0.011	0.037***	0.028**	-0.155	0.011	-0.003
	(0.012)	(0.012)	(0.008)	(0.013)	(0.014)	(0.192)	(0.011)	(0.012)
Size > 60B	-0.043***	0.053***	-0.009	-0.018	0.026*	-0.142	0.028***	0.034***
	(0.011)	(0.011)	(0.009)	(0.013)	(0.015)	(0.139)	(0.010)	(0.011)
Open-source	-0.011	0.029**	0.034***	0.009	-0.053***	-0.163	0.001	0.002
	(0.014)	(0.014)	(0.010)	(0.016)	(0.017)	(0.118)	(0.012)	(0.013)
Investor characteristics								
Age = 45	-0.050***	0.050***	-0.001	0.004	0.006	-0.130	-0.012	-0.013
	(0.013)	(0.013)	(0.011)	(0.015)	(0.017)	(0.086)	(0.012)	(0.013)
Age = 60	-0.096***	0.084***	0.012	0.033**	0.034**	0.112	-0.071***	-0.079***
	(0.012)	(0.012)	(0.009)	(0.014)	(0.015)	(0.214)	(0.011)	(0.012)
Risk tolerance = high	0.301***	-0.289***	-0.044***	-0.007	0.005	1.378***	0.256***	0.113***
	(0.011)	(0.010)	(0.008)	(0.012)	(0.013)	(0.145)	(0.010)	(0.011)
Home country = US	-0.009	0.009	-0.006	-0.010		0.060	-0.001	-0.010
	(0.014)	(0.013)	(0.006)	(0.015)		(0.072)	(0.013)	(0.013)
Home country = Germany	-0.016	0.019	0.033***	-0.061***	-0.698***	0.500	0.002	0.027
	(0.019)	(0.017)	(0.010)	(0.020)	(0.015)	(0.478)	(0.017)	(0.017)
Home country = China	0.020	-0.030*	-0.244***	-0.167***	-0.481***	0.458***	-0.054***	-0.146***
	(0.018)	(0.017)	(0.019)	(0.022)	(0.021)	(0.160)	(0.017)	(0.019)
Sustainability = Yes	0.093***	-0.083***	0.003	0.006	-0.019	0.342*	0.011	-0.011
	(0.012)	(0.011)	(0.009)	(0.013)	(0.015)	(0.177)	(0.011)	(0.012)
Data available	-0.067	0.082**	0.279***	0.464***	-0.358***	0.605	0.102***	0.156***
	(0.042)	(0.039)	(0.040)	(0.039)	(0.044)	(0.545)	(0.034)	(0.039)
Some response error	-0.034***	0.002	-0.023**	0.005	0.084***	-0.075	-0.070***	-0.057***
	(0.012)	(0.012)	(0.010)	(0.013)	(0.015)	(0.177)	(0.011)	(0.012)
Constant	0.650***	0.297***	0.664***	0.233***	1.041***	1.581***	0.321***	0.577***
	(0.047)	(0.043)	(0.041)	(0.045)	(0.049)	(0.580)	(0.039)	(0.044)
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	0.050	0.360	0.196

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 model type, "$\leq$ 60B" for model 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 B8: Determinants of portfolio performance and fees (robustness)

	(1) Mean monthly return (%)	(2) Mean monthly return (%)	(3) Sharpe ratio	(4) Sharpe ratio	(5) FF6 α	(6) FF6 α	(7) Total exp. ratio (%)	(8) Total exp. ratio (%)
Model characteristics								
Foundation model	-0.017	-0.021	-0.026*	-0.044**	0.019	0.030	-0.008	0.015
	(0.020)	(0.024)	(0.015)	(0.018)	(0.017)	(0.020)	(0.011)	(0.015)
Size > 60B	-0.015	-0.021	-0.015	-0.032**	-0.011	-0.004	-0.023**	0.002
	(0.014)	(0.017)	(0.012)	(0.013)	(0.012)	(0.014)	(0.009)	(0.009)
Open-source	-0.049***	-0.026	-0.036**	-0.007	-0.057***	-0.025	-0.021*	0.011
	(0.018)	(0.022)	(0.017)	(0.019)	(0.018)	(0.022)	(0.011)	(0.012)
Data available	0.273***	0.320***	0.292***	0.311***	0.127***	0.134***	-0.247***	-0.191***
	(0.052)	(0.069)	(0.046)	(0.054)	(0.036)	(0.049)	(0.060)	(0.068)
Some response error	0.005	0.017	-0.028*	-0.018	0.040**	0.044**	0.057***	0.046***
	(0.019)	(0.024)	(0.015)	(0.016)	(0.016)	(0.018)	(0.010)	(0.011)
Cutoff within 6 months		0.005		-0.048***		0.027*		0.040***
		(0.015)		(0.017)		(0.016)		(0.013)
Constant	-0.138**	-0.214***	-0.126**	-0.129**	-0.298***	-0.356***	0.373***	0.243***
	(0.057)	(0.081)	(0.050)	(0.061)	(0.041)	(0.062)	(0.062)	(0.065)
Obs.	1,559	1,186	1,559	1,186	1,559	1,186	1,534	1,166
Adj. R2	0.121	0.114	0.191	0.205	0.062	0.059	0.158	0.125
Profile controls	✓	✓	✓	✓	✓	✓	✓	✓

Note: Omitted categories are "fine-tuned" for model type, "$\leq$ 60B" for model 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 B9: Summary statistics for alternative factor model specifications

	N	Mean	Median	SD	Min	Max
Panel A: Developed market factors, monthly rebalancing						
Mean monthly return (%)	1,559	0.19	0.15	0.33	-0	8
Monthly volatility (%)	1,559	2.88	2.78	3.14	0	114
Annual Sharpe ratio	1,559	0.16	0.20	0.28	-3	1
FF6 alpha (%)	1,559	-0.11	-0.16	0.26	-1	4
FF6 market beta	1,559	0.51	0.51	0.24	0	1
Idiosyncratic volatility (%)	1,559	0.71	0.79	0.23	-0	1
Panel B: Region-specific factors, monthly rebalancing						
Mean monthly return (%)	1,559	0.19	0.15	0.33	-0	8
Monthly volatility (%)	1,559	2.88	2.78	3.14	0	114
Annual Sharpe ratio	1,559	0.16	0.20	0.28	-3	1
FF6 alpha (%)	1,559	-0.08	-0.14	0.31	-1	6
FF6 market beta	1,559	0.49	0.49	0.23	0	1
Idiosyncratic volatility (%)	1,559	0.70	0.75	0.22	-0	1
Panel C: Region-specific factors, annual rebalancing						
Mean monthly return (%)	1,559	0.18	0.13	0.39	-1	11
Monthly volatility (%)	1,559	2.96	2.67	4.91	0	182
Annual Sharpe ratio	1,559	0.16	0.17	0.26	-3	1
FF6 alpha (%)	1,559	-0.11	-0.16	0.31	-1	6
FF6 market beta	1,559	0.48	0.49	0.21	-2	1
Idiosyncratic volatility (%)	1,559	0.70	0.76	0.20	-0	1

Note: Summary statistics are displayed on the model-profile level (32 models × 64 profiles). Panel A is the main specification (see Table 3) 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 price data from Bloomberg.

Table B10: Determinants of portfolio risk exposure (region-specific factor portfolios)

	Monthly rebalancing			Annual rebalancing		
	(1) Monthly volatility (%)	(2) FF6 β_{mkt}	(3) Idiosyncratic volatility	(4) Monthly volatility (%)	(5) FF6 β_{mkt}	(6) Idiosyncratic volatility
Model characteristics						
Foundation model	-0.159	0.011	-0.004	-0.338	0.010	-0.013
	(0.200)	(0.011)	(0.011)	(0.325)	(0.011)	(0.011)
Size > 60B	-0.145	0.029***	0.042***	-0.240	0.038***	0.049***
	(0.147)	(0.009)	(0.010)	(0.231)	(0.009)	(0.009)
Open-source	-0.137	-0.016	-0.021*	-0.166	-0.004	-0.018
	(0.088)	(0.012)	(0.012)	(0.141)	(0.011)	(0.011)
Investor characteristics						
Age = 45	-0.137	-0.012	-0.009	-0.154	-0.007	-0.008
	(0.087)	(0.012)	(0.011)	(0.116)	(0.011)	(0.011)
Age = 60	0.100	-0.068***	-0.082***	0.396	-0.051***	-0.068***
	(0.206)	(0.011)	(0.011)	(0.330)	(0.011)	(0.011)
Risk tolerance = high	1.389***	0.250***	0.117***	1.227***	0.197***	0.076***
	(0.154)	(0.009)	(0.010)	(0.245)	(0.009)	(0.009)
Home country = US	0.069	0.003	-0.007	-0.010	-0.014	-0.010
	(0.071)	(0.013)	(0.013)	(0.087)	(0.012)	(0.012)
Home country = Germany	0.362	-0.030*	-0.014	0.608	-0.027*	-0.022
	(0.396)	(0.016)	(0.016)	(0.632)	(0.015)	(0.015)
Home country = China	0.372***	-0.163***	-0.249***	0.632***	-0.154***	-0.224***
	(0.144)	(0.015)	(0.014)	(0.234)	(0.015)	(0.014)
Sustainability = Yes	0.338*	0.004	-0.022**	0.500	0.011	-0.030***
	(0.195)	(0.010)	(0.010)	(0.310)	(0.010)	(0.010)
Constant	2.144***	0.414***	0.727***	2.264***	0.414***	0.738***
	(0.136)	(0.019)	(0.020)	(0.204)	(0.019)	(0.019)
Obs.	1,559	1,559	1,559	1,559	1,559	1,559
Adj. R2	0.050	0.376	0.279	0.019	0.296	0.234

Note: The table reports regression coefficients of risk measures to model 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 B11: Dominance statistics for exposure regressions (region-specific factor portfolios)

	Monthly rebalancing						Annual rebalancing					
	(1)			(2)			(3)			(4)		
	FF6 β_{mkt}			Idiosyncratic volatility			FF6 β_{mkt}			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, 0.01]	0.00	0.00	[0.00, 0.00]	0.00	0.00	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]
Size ¿ 60B	0.00	0.01	[0.00, 0.01]	0.01	0.03	[0.00, 0.02]	0.01	0.03	[0.00, 0.02]	0.01	0.06	[0.01, 0.03]
Open-source	0.00	0.01	[0.00, 0.01]	0.00	0.01	[0.00, 0.01]	0.00	0.00	[0.00, 0.01]	0.00	0.02	[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	0.04	[0.01, 0.03]	0.03	0.09	[0.01, 0.04]	0.01	0.04	[0.00, 0.02]	0.02	0.08	[0.01, 0.03]
Risk tolerance = high	0.29	0.76	[0.25, 0.33]	0.07	0.24	[0.05, 0.09]	0.22	0.72	[0.18, 0.25]	0.03	0.14	[0.02, 0.05]
Home country = US	0.01	0.02	[0.00, 0.02]	0.02	0.06	[0.01, 0.03]	0.01	0.02	[0.00, 0.01]	0.02	0.07	[0.01, 0.03]
Home country = Germany	0.00	0.00	[0.00, 0.00]	0.01	0.02	[0.00, 0.01]	0.00	0.01	[0.00, 0.00]	0.01	0.02	[0.00, 0.01]
Home country = China	0.05	0.14	[0.04, 0.07]	0.15	0.51	[0.12, 0.18]	0.05	0.18	[0.04, 0.07]	0.14	0.57	[0.11, 0.17]
Sustainability = yes	0.00	0.00	[0.00, 0.00]	0.00	0.02	[0.00, 0.01]	0.00	0.00	[0.00, 0.00]	0.01	0.03	[0.00, 0.02]
Obs.		1,559			1,559			1,559			1,559	
R2		0.38			0.28			0.30			0.24	

Note: The table reports standardized dominance statistics for regressions of six-factor market beta (columns 1 and 3) and idiosyncractic 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 B12: Impact of risk tolerance on equity share and portfolio risk, by model type and size (region-specific factor portfolios)

	Monthly rebalancing				Annual rebalancing			
	(1) FF6 β_{mkt}	(2) FF6 β_{mkt}	(3) Idiosyncratic volatility	(4) Idiosyncratic volatility	(5) FF6 β_{mkt}	(6) FF6 β_{mkt}	(7) Idiosyncratic volatility	(8) Idiosyncratic volatility
Risk tolerance = high	0.210***	0.225***	0.083***	0.105***	0.166***	0.175***	0.052***	0.067***
	(0.014)	(0.013)	(0.015)	(0.015)	(0.015)	(0.013)	(0.014)	(0.014)
Model type								
Risk tolerance = high × original model	0.066***		0.056***		0.050***		0.040**	
	(0.019)		(0.019)		(0.018)		(0.018)	
Model size								
Risk tolerance = high × Size > 60B		0.049***		0.022		0.043**		0.016
		(0.019)		(0.019)		(0.018)		(0.018)
Constant	0.434***	0.425***	0.744***	0.732***	0.430***	0.424***	0.750***	0.741***
	(0.020)	(0.019)	(0.021)	(0.020)	(0.019)	(0.019)	(0.020)	(0.019)
Obs.	1,559	1,559	1,559	1,559	1,559	1,559	1,559	1,559
Adj. R2	0.381	0.379	0.283	0.280	0.299	0.298	0.236	0.234
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-model (vs. proprietary), and three dummy variables indicating model size: ≤60B indicates models with more than 7B, but at most 60B parameters. ≤75B indicates models with more than 60B, but at most 75B parameters. ≥75B indicates models with more than 75B parameters. Models with at most 7B parameters are the baseline group. Heteroskedasticity-robust standard errors are reported in parentheses ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

Table B13: Determinants of portfolio performance and fees (region-specific factor portfolios)

	Monthly rebalancing		Annual rebalancing	
	(1)	(2)	(3)	(4)
	FF6	FF6	FF6	FF6
	α_{mkt}	α_{mkt}	α_{mkt}	α_{mkt}
Model characteristics				
Foundation model	0.019	0.034	0.010	0.038
	(0.020)	(0.024)	(0.022)	(0.024)
Size > 60B	-0.023*	-0.018	-0.024*	-0.009
	(0.014)	(0.015)	(0.013)	(0.015)
Open-source	-0.036*	-0.011	-0.036	-0.020
	(0.019)	(0.024)	(0.023)	(0.030)
Cutoff within 6 months		0.035*		0.054***
		(0.018)		(0.018)
Investor characteristics				
Age = 45	0.004	0.014	0.006	0.013
	(0.014)	(0.016)	(0.014)	(0.016)
Age = 60	0.054***	0.070***	0.070***	0.080***
	(0.019)	(0.023)	(0.019)	(0.023)
Risk tolerance = high	0.092***	0.091***	0.070***	0.073***
	(0.014)	(0.017)	(0.015)	(0.018)
Home country = US	0.029**	0.024	0.027**	0.025
	(0.012)	(0.015)	(0.013)	(0.016)
Home country = Germany	-0.002	-0.014	0.009	-0.004
	(0.017)	(0.019)	(0.017)	(0.019)
Home country = China	0.238***	0.236***	0.257***	0.259***
	(0.033)	(0.041)	(0.033)	(0.041)
Sustainability = Yes	-0.012	-0.024	-0.000	-0.005
	(0.017)	(0.020)	(0.017)	(0.020)
Constant	-0.175***	-0.229***	-0.198***	-0.272***
	(0.029)	(0.047)	(0.030)	(0.045)
Obs.	1,559	1,186	1,559	1,186
Adj. R2	0.103	0.098	0.108	0.107

Note: Omitted categories are "fine-tuned" for model type, "≤ 60B" for model 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 B14: Home bias: heterogeneity

	All countries			China	Germany	United States
	(1) $d_i - m_i$	(2) $\frac{d_i - m_i}{m_i}$	(3) $\frac{d_i - m_i}{1 - m_i}$	(4) $d_i - m_i$	(5) $d_i - m_i$	(6) $d_i - m_i$
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 model and profile characteristics. Columns 1 through 3 pools 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 parantheses ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).

Table B15: Gender-based discrimination: heterogeneity

	(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	0.056	0.043	-0.026
	(0.052)	(0.021)	(0.040)	(0.041)	(0.022)
Size > 60B	-0.046	0.012	-0.027	0.006	0.023
	(0.044)	(0.020)	(0.034)	(0.044)	(0.015)
Open-source	0.016	0.003	0.043	0.139***	-0.021
	(0.049)	(0.022)	(0.045)	(0.051)	(0.020)
Constant	0.027	0.001	-0.029	-0.105	0.014
	(0.071)	(0.033)	(0.054)	(0.064)	(0.018)
Obs.	88	88	88	88	128
Adj. R2	-0.014	-0.025	0.006	0.071	0.005

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

Table B16: 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 model and profile characteristics. Heteroskedasticity-robust standard errors are reported in parantheses ($^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$).

Table B17: 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 model and profile characteristics. Heteroskedasticity-robust standard errors are reported in parantheses ($^{*}p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$).