Biased Echoes: Generative AI Models Reinforce Investment Biases and Increase Portfolio Risks of Private Investors

Philipp Winder[1], Christian Hildebrand[1,*], and Jochen Hartmann[2]

[1] University of St.Gallen, Institute of Behavioral Science & Technology, St. Gallen, 9000, Switzerland
[2] Technical University of Munich, TUM School of Management, Munich, 80333, Germany
[*] christian.hildebrand@unisg.ch

ABSTRACT

Generative AI models are increasingly used by private investors seeking financial advice. The current paper examines the potential of these models to perpetuate investment biases and affect the economic security of individuals at scale. It provides a systematic assessment of how generative AI models used for investment advice shape the portfolio risks of private investors. We offer a comprehensive model of generative AI investment advice risk, examining five key dimensions of portfolio risks (geographical cluster risk, sector cluster risk, trend chasing risk, active investment allocation risk, and total expense risk). We demonstrate across four studies that generative AI models used for investment advice induce increased portfolio risks across all five risk dimensions, and that a range of debiasing interventions only partially mitigate these risks. Our findings show that generative AI models exhibit similar “cognitive” biases as human investors, reinforcing existing investment biases inherent in their training data.

Keywords: generative AI, large language models, private investors, retail investors, financial portfolio risks, financial decision making

Introduction

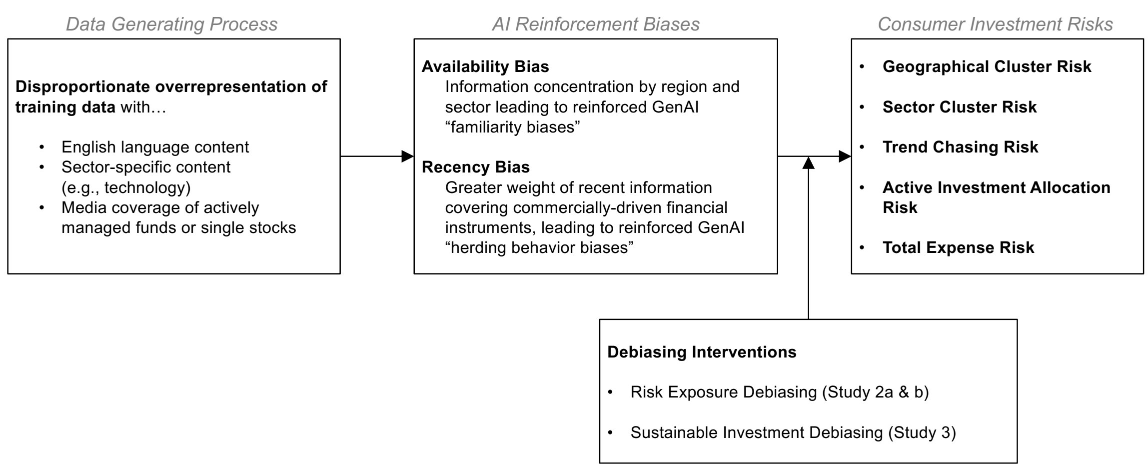
The contemporary landscape of financial advisory is undergoing a paradigm shift, with millions of private investors increasingly relying on AI-powered advisory services [1–3]. A new class of such AI-based services is emerging, which uses Generative AI (GenAI) models such as OpenAI's ChatGPT to offer financial advice to private investors [4]. In a recent study, 20% of UK private investors with over £10,000 invested used ChatGPT for financial advice, with 73% believing that ChatGPT can provide reliable financial advice [4]. The current research examines whether such GenAI financial advice is truly reliable and to which extent these recommendations may carry unexplored and disproportionate risks for private investors. With millions of users starting to employ generative AI systems to receive financial advice, believing that large language models and AI systems can offer sound financial advice has the potential to significantly shape (and induce) not only idiosyncratic risks for single individuals but systemic investment risks across financial markets at scale[5].

We predict that GenAI models can lead to systematic investment risks due to the underlying training data: The overrepresentation of English language corpora [6] and therefore biasing advice toward English-speaking (especially North-American) equities; and the overrepresentation of themes or events covered by publicly available text corpora [6–9] and therefore biasing advice toward overrepresented sectors (such as technology or consumer staples) and chasing recent trends (see Fig. 1). We propose a comprehensive framework of risk evaluation that explains how these inherent training data biases lead to (1.) greater asset concentration (geographical and sector cluster risk), (2.) a riskier equity structure of a portfolio (active asset allocation and total expense risk), and (3.) riskier time-dependent trading decisions (trend chasing risk).

We expect a high concentration of US-based investments (i.e. geographical cluster risk) in GenAI financial advice due to the training data stemming especially from English language content and the fact that the US is one of the largest and most developed economies in the world with a strong presence in media reports. Similarly, we propose that due to the strong media coverage of, for example, sectors such as technology or consumer staples compared to other sectors that are of similar or even greater economic weight in terms of contribution to the GDP (such as transportation or service sectors [10,11]) lead to an over-investment in such sectors (i.e. sector cluster risk). We hypothesize that GenAI financial advice, due to the inherent recency bias, may respond more strongly to recent events than passive indices (i.e. trend chasing risk), potentially leading to poorly timed trades [12]. This could amplify market bubbles and market volatility [13]. We further propose that GenAI might favor actively managed assets (i.e. active asset allocation risk), due to the greater likelihood of coverage of high-profile stocks [14] with a subsequent increase in the overall costs of investment (i.e. total expense risk) [15,16].

In four large-scale studies prompting three major GenAI systems (OpenAI's ChatGPT, Google's Gemini, and Microsoft's Copilot), we observe a systematic increase in portfolio risks across all five dimensions for private investors. We find that both narrow debiasing prompts (such as directly stating to avoid any management fees) and broader debiasing prompts (such as more broadly stating to avoid active management and ensure high diversification) only partially mitigate these risks. Even highly specific debiasing prompt interventions such as directly asking to ensure a sustainable, globally diversified portfolio as a "socially responsible private investor" only moderately reduces the extent of investment risk. Our findings show that GenAI models exhibit similar "cognitive" biases as human investors, reinforcing existing investment biases due to the nature of their training data.

Figure 1. GenAI Investment Risk Model



Results

GenAIs Increase All Five Portfolio Investment Risk Types for Private Investors

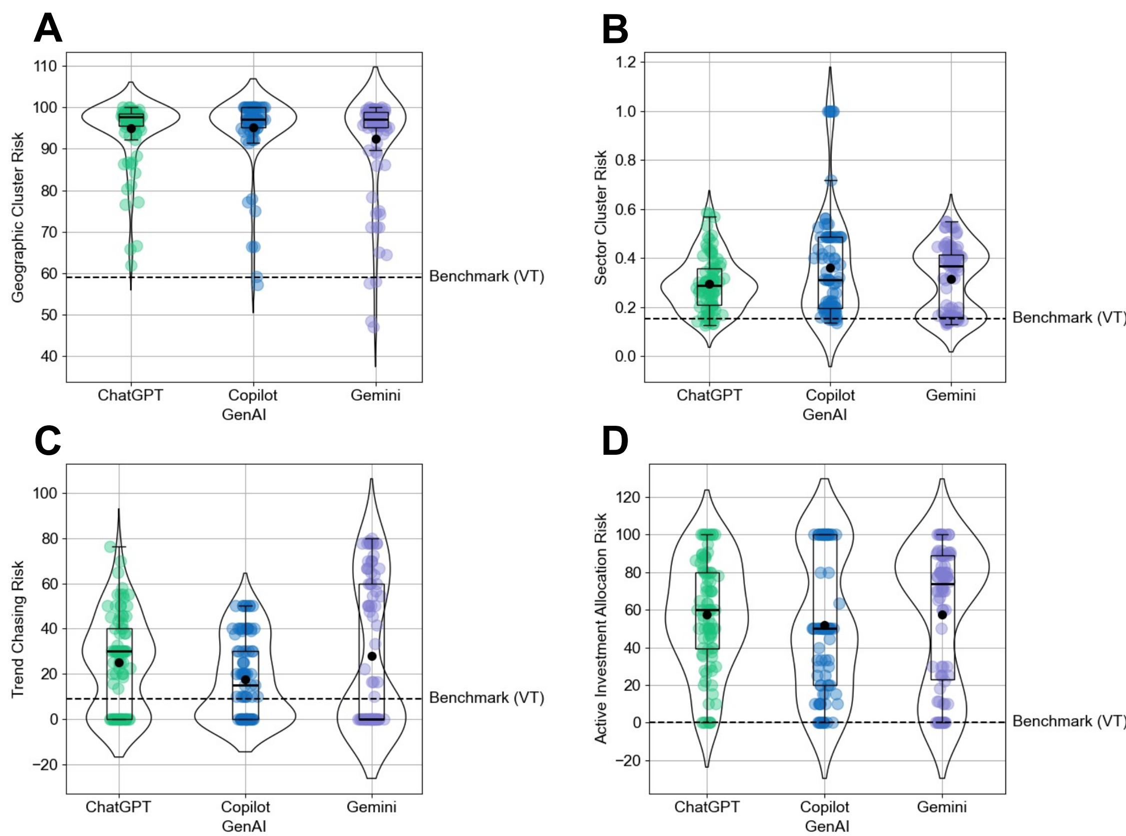
In Study 1 we assessed the type of financial advice that private investors receive as a function of their appetite for risk (self-stated risk-taking tendency and demographic-related risks such as investors' age) from the most widely used GenAIs (Open AI's ChatGPT, Google's Gemini, Microsoft's Copilot). Study 1 employed an experimental paradigm using a 3 (risk tendency: high / medium / low) x 3 (age: 15 / 30 / 50) x 3 (GenAI: ChatGPT / Copilot / Gemini) x 10 (number of queries per experimental condition) full factorial design (see Methods section for details). Each experimental condition was run ten times per GenAI and averaged across these batches. The resulting financial advice was augmented by using Yahoo Finance and the Refinitiv Eikon data, to compute our five risk measures of interest (geographical cluster risk, sector cluster risk, trend chasing risk, active investment allocation risk, and total expense risk) and a comparison with a passive ETF (Exchange Traded Fund) benchmark followed (see Methods section).

Geographical & Sector Cluster Risk. All three GenAIs revealed excessive cluster risk compared to our benchmark index that is reflected in both an over-investment in US stocks (Fig. 2; Panel A; $M_{GeminiUS}$= .9249, $SD_{GeminiUS}$ = .1195; $M_{CopilotUS}$= .9514, $SD_{CopilotUS}$ = .0842; $M_{ChatGPTUS}$= .9488, $SD_{ChatGPTUS}$ = .0746; $M_{Benchmark}$ = .5896; all p's < .001) and a higher concentration of funds in individual sectors (Fig 2; Panel B; M_{Gemini}= .3137, SD_{Gemini} = .1371; $M_{Copilot}$= .3621, $SD_{Copilot}$ = .221; $M_{ChatGPT}$= .2957, $SD_{ChatGPT}$ = .1103; $M_{Benchmark}$= .1528; all p's < .001).

Trend Chasing Risk. As shown in Fig. 2 Panel C, we find that all GenAIs heavily engage in trend chasing with up to 27.92% invested in the top three equities that were traded most frequently in the past three months prior to prompting the GenAIs. Consistent with our predictions, we find that all three GenAIs display systematically higher trend chasing compared to the benchmark index (M_{Gemini} = .2792, SD_{Gemini} = .3223; $M_{Copilot}$ = .1761, $SD_{Copilot}$ = .1779; $M_{ChatGPT}$ = .2492, $SD_{ChatGPT}$= .2061; $M_{Benchmark}$ = .09; all p's < .001).

Active Investment Allocation Risk. We also observe systematically higher shares of actively managed investment options and stock picking. Specifically, we find that over 50% of investments were allocated into actively managed funds or single equities (Fig. 2; Panel D; M_{Gemini} = .5747, SD_{Gemini} = .3655; $M_{Copilot}$ = .5185, $SD_{Copilot}$ = .3662; $M_{ChatGPT}$ = .5741, $SD_{ChatGPT}$ = .2921; $M_{Benchmark}$ = 0; all *p's* < .001).

Figure 2. GenAI financial advice increases financial investment portfolio risks (Study 1).



Total Expense Risk. Finally, in line with our proposition, we observe significantly higher total expense ratios (TER) across all GenAI portfolio recommendations compared to the benchmark index (M_{Gemini}= .1537%, SD_{Gemini}= .2319; $M_{Copilot}$= .1265%, $SD_{Copilot}$= .1768; $M_{ChatGPT}$= .2013%, $SD_{ChatGPT}$= .2112; $M_{Benchmark}$= .07%; $p_{Gemini\text{-}Benchmark}$ < .01; $p_{Copilot\text{-}Benchmark}$ = .011; $p_{ChatGPT\text{-}Benchmark}$ < .001).

Ancillary Findings: Portfolio Returns. As summarized in Appendix C, we performed a detailed financial performance analysis of the six-month period *after* having received the advice relative to our benchmark index (July 1st, 2023 – January 1st, 2024). We find no difference relative to our benchmark for

ChatGPT and Copilot for the unadjusted returns and only a slight overperformance for Gemini ($M_{ChatGPT}$ = .0655, $SD_{ChatGPT}$ = .0413; $M_{Copilot}$ = .0724, $SD_{Copilot}$ = .0566; M_{Gemini} = .0832, SD_{Gemini} = .041; $M_{Benchmark}$ = .0703; $p_{Gemini\text{-}Benchmark} < .01$; $p_{Copilot\text{-}Benchmark} = .724$; $p_{ChatGPT\text{-}Benchmark} = .269$). However, we find a systematically *lower* risk-adjusted performance (or almost equal performance in the case of Gemini) relative to the benchmark, due to the greater volatility and risk concentration shown in the preceding analyses ($M_{ChatGPT}$ = 2.03, $SD_{ChatGPT}$ = 3.97; $M_{Copilot}$ = -10.13, $SD_{Copilot}$ = 41.63; M_{Gemini} = 3.72, SD_{Gemini} = 3.42; $M_{Benchmark}$ = 3.62; $p_{Gemini\text{-}Benchmark} = .78$; $p_{Copilot\text{-}Benchmark} < .01$; $p_{ChatGPT\text{-}Benchmark} < .001$).

Ancillary Findings: Language Style. Finally, we examined the specific language style employed across all studies (see Appendix B for details). We utilized a zero-shot transformer model (Facebook's BART-large trained on the MultiNLI dataset; [17]) to detect whether the investment advice offered a clear rationale (i.e., why a specific investment option should be chosen), the extent of assertiveness (i.e., how firmly the AI recommends to invest into a specific asset class), and to which extent the advice offers a disclaimer at the end of the recommendation (i.e., stating the potential risks involved with the investment). These exploratory analyses demonstrate that all recommendations offered a seemingly plausible explanation (e.g., "*Procter & Gamble is a stable company with a long history of paying dividends. Dividend stocks provide regular income and can grow over time*."), with medium to high assertiveness (e.g., "Considering your requirements, here's a diversified investment portfolio with a breakdown of how much you *should allocate* to each type of investment:"), and offering a seemingly trustworthy and "caring" disclaimer in the recommendation (e.g., "Remember, these are *just recommendations*, and it's crucial to do your own research or consult a financial advisor before making any investment decisions.").

Narrow Debiasing Interventions Only Partially Mitigate Portfolio Risks for Private Investors

In Study 2a we assessed whether we can alter a single risk dimension (such as avoiding management fees) without changing the other types of risks. We used a two-cell experimental design (control prompt vs. debiasing intervention prompt) with a control prompt identical to Study 1 and a debiasing intervention prompt that explicitly requested no management fees ("*I don't want to pay any* *management fees.*”). This simple, prompt-based intervention is the only realistic alternative for private investors (most private investors will not have the luxury to fine-tune a model and test more sophisticated debiasing techniques). The baseline paradigm in both the control condition and risk debiasing condition was identical and used the same experimental setup as in Study 1 (full factorial design with 3 (risk tendency: high / medium / low) x 3 (age: 15 / 30 / 50) x 2 (condition: control prompt vs. debiasing intervention prompt) x 10 (number of queries per experimental prompt configuration); given the lack of differences between the three GenAI systems, we focused on ChatGPT as the largest commercially available GenAI system for the remainder of the paper).

Total Expense Risk. Narrow debiasing intervention prompts only directionally reduce the TER compared to the control condition ($M_{Control}$= .1908%, $SD_{Control}$= .2099%; $M_{Debiased}$= .1759%, $SD_{Debiased}$= .2006; $M_{Benchmark}$ = .07%; $p_{Control\text{-}Debiased}$ = .645 all other *p’s* < .001).

Active Investment Allocation Risk. We find that the debiasing intervention moderately reduces the share of actively managed investments. Yet, we still find that over 48% of investments are allocated to actively managed funds or single equities ($M_{Control}$= .5957, $SD_{Control}$= .3099; $M_{Debiased}$= .4872, $SD_{Debiased}$= .3964; $M_{Benchmark}$ = 0; $p_{Control\text{-}Debiased}$ = .042; all other *p’s* < .001).

Geographical & Sector Cluster Risk. A narrower debiasing intervention prompt also reduced the over-investment in US stocks ($M_{Control}$= .9494, $SD_{Control}$= .0889; $M_{Debiased}$= .9039, $SD_{Debiased}$= .1169; $M_{Benchmark}$ = .5896; $p_{Control\text{-}Debiased}$ < .01; all other *p’s* < .001) and sector concentration ($M_{Control}$= .3201, $SD_{Control}$= .1332; $M_{Debiased}$= .277, $SD_{Debiased}$= .1411; $M_{Benchmark}$ = .1528; $p_{Control\text{-}Debiased}$ = .037; all other *p’s* < .001) in GenAI financial advice. However, the geographical as well as cluster risk remained significantly larger compared to the benchmark.

Trend Chasing Risk. The narrow debiasing intervention prompt did not reduce trend chasing, resulting in still significantly stronger trend chasing compared to the benchmark index ($M_{control}$= .245, $SD_{Control}$= .2003; $M_{Debiased}$= .2364, $SD_{Debiased}$= .2426; $M_{Benchmark}$ = .09; $p_{Control\text{-}Debiased}$ = .796; all other *p’s* < .001).

Broader Debiasing Interventions are More Effective in Mitigating Portfolio Risks

In Study 2b we assessed to which extent a broader debiasing prompt may reduce the overall financial portfolio risk. Specifying multiple risks in the debiasing prompt, this prompting strategy may reinforce the risk reduction of every single risk alone as such broad debiasing interventions can often lead to effective debiasing or risk reduction GenAI models [18]. We tested this possibility by employing the same experimental design as in Study 2a, with an overarching debiasing intervention in the prompt ("Avoid common investment mistakes such as lack of diversification, cluster risks, and active management").

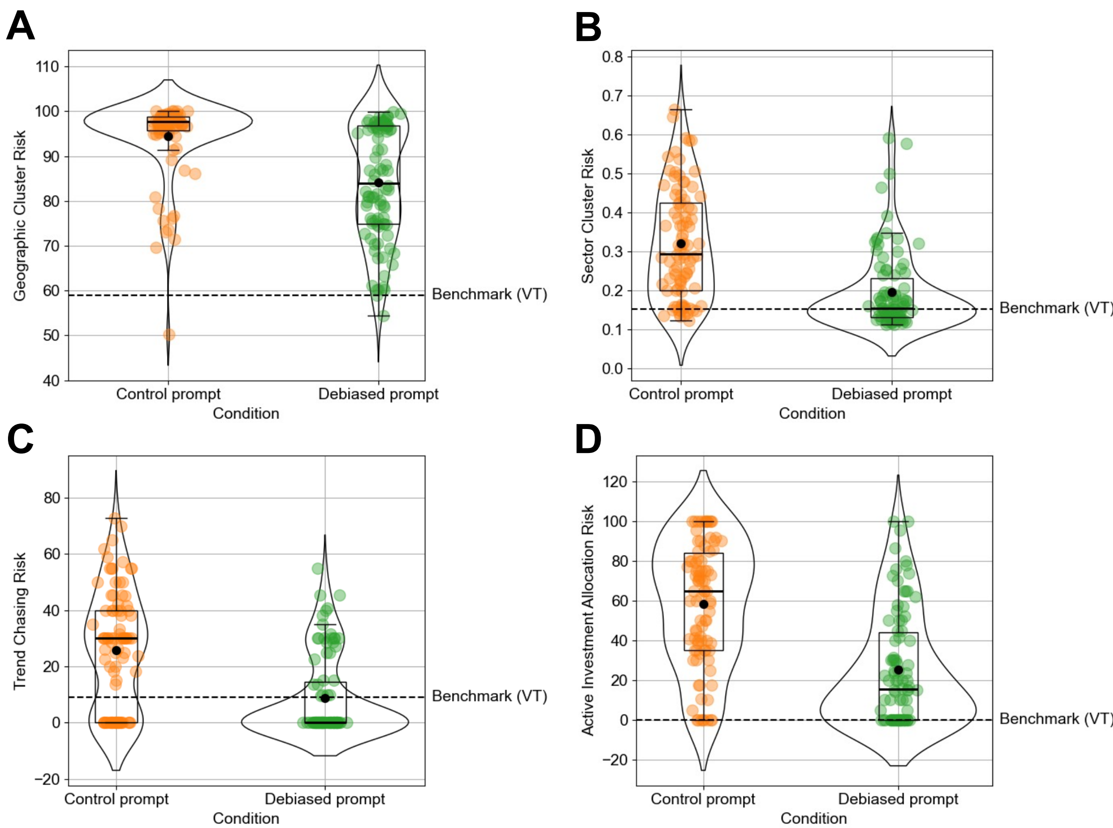
Geographical & Sector Cluster Risk. While the debiasing prompt reduced the over-investment in US assets (Fig. 3; Panel A; $M_{Control}$= .9453, $SD_{Control}$= .0861; $M_{Debiased}$= .8406, $SD_{Debiased}$= .1271; $M_{Benchmark}$ = .5896; all *p's* < .001) and sector concentration (Fig. 3; Panel B; $M_{Control}$= .3201, $SD_{Control}$= .1405; $M_{Debiased}$= .196, $SD_{Debiased}$= .0995; $M_{Benchmark}$ = .1528; all *p's* < .001), both risks were still significantly greater compared to the benchmark. These findings highlight that a more equal distribution across sectors is easier to achieve compared to an equal distribution across financial markets (i.e., arguably due to the strong presence of US-based equities).

Trend Chasing Risk. We find that the debiasing intervention significantly reduced trend chasing (Fig. 3; Panel C; $M_{Control}$= .2587, $SD_{Control}$= .2089; $M_{Debiased}$= .0884, $SD_{Debiased}$= .1451; $F(1, 178) = 40.35$; $p < .001$), to the extent that it no longer differs significantly from the benchmark ($M_{Debiased}$= .0884, $SD_{Debiased}$= .1451; $M_{Benchmark}$ = .09; $t(89) = -.11$; $p = .915$).

Active Investment Allocation Risk. As shown in Fig. 3; Panel D, we find that the debiasing intervention reduces the share of actively managed investment options by over 32% but fails to reduce it to the level of the benchmark ($M_{Control}$= .5812, $SD_{Control}$= .3125; $M_{Debiased}$= .2519, $SD_{Debiased}$= .2843; $M_{Benchmark}$ = 0; all *p's* < .001)

Total Expense Risk. Finally, the debiasing intervention significantly decreased TER by more than .06%. Nevertheless, TER remained notably higher compared to our benchmark ($M_{Control}$= .2027%, $SD_{Control}$= .212%; $M_{Debiased}$= .137%, $SD_{Debiased}$= .1406%; $M_{Benchmark}$ = .07%; $p_{Control-Debiased}$ = .016; all other *p's* < .001).

Figure 3. Broad debiasing reduces the overall financial investment portfolio risks in GenAI financial advice (Study 2b).



Incorporating Specific Investment Goals to Partially Mitigate Portfolio Risks

In Study 3 we tested whether incorporating an explicit investment goal in the prompt further aids in mitigating investment portfolio risks in GenAI investment advice. Thus, the study directly tests the ability of GenAI systems to adapt their investment advice based on contextual information. Instead of directly asking a GenAI system to diversify a portfolio, a private investor may prompt a specific investment goal that would in turn translate into a more diversified portfolio merely based on the contextual information provided. The current study tests this possibility to correctly infer and "sense" such context information. Again, we employed exactly the same experimental design as in Study 2a but incorporated an explicit investment goal in the prompt ("*I want to invest in a way that promotes responsible and socially impactful contributions to our global society.*").

Geographical & Sector Cluster Risk. We observe a significant decrease in over-investments in US Stocks in the investment goal prompt compared to control, yet higher than our benchmark ($M_{Control}$= .9518, $SD_{Control}$= .071; M_{Goal}= .8905, SD_{Goal}= .1344; $M_{Benchmark}$ = .5896; all *p's* < .001). However, we find that the introduction of a socially responsible investment goal did not lead to a significant decrease in sector concentration compared to the control prompt and remains significantly higher than the benchmark index ($M_{Control}$= .3331, $SD_{Control}$= .146; M_{Goal}= .3386, SD_{Goal}= .1938; $M_{Benchmark}$ = .1528; $p_{Control\text{-}Goal}$ = .831; all other *p's* < .001). As shown in Fig. 4; Panel A, not only remains this sector concentration very high, but is also substantially overleveraged towards utilities, due to shifting investments towards the energy sector.

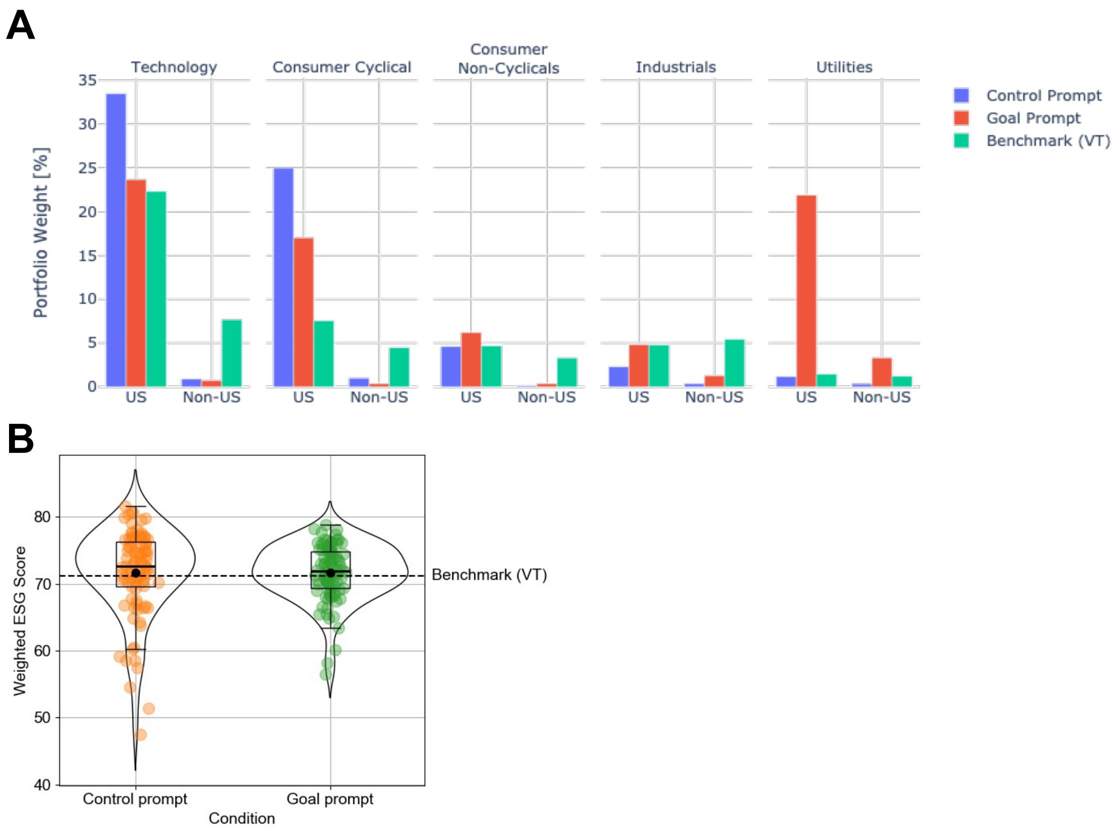
Next, we further examined the ESG score in the intervention (vs. control) condition. As summarized in Fig. 4; Panel B, specific investment goals were not effective in increasing the portfolio ESG rating of the portfolio. Specifically, we find that there is no significant portfolio ESG rating difference when incorporating a social responsibility goal prompt compared to the control prompt or benchmark ($M_{Control}$= 71.59, $SD_{Control}$= 6.6; M_{Goal}= 71.61, SD_{Goal}= 4.25; $M_{Benchmark}$ = 71.12; $p_{Control\text{-}Goal}$ = .982; $p_{Control\text{-}Benchmark}$ = .504, $p_{Goal\text{-}Benchmark}$ = .282). However, incorporating the social responsibility goal effectively reduced the risk of receiving a low ESG rating compared to the control ($F(1, 693) = 23.83$, $p < .001$).

Trend Chasing Risk. We also observe a significant decrease in trend chasing behavior in GenAI investment advice when incorporating a social responsibility investment goal ($M_{Control}$= .2695, $SD_{Control}$= .2246; M_{Goal}= .0763, SD_{Goal}= .1053; $M_{Benchmark}$ = .09; $p_{Goal\text{-}Benchmark}$ = .221; all other *p's* < .001).

Active Investment Allocation Risk. Adding the social responsibility investment goal however, increased the actively managed portfolio portion by over 29% compared to the control condition ($M_{Control}$= .5941, $SD_{Control}$= .3045; M_{Goal}= .8851, SD_{Goal}= .1739; $M_{Benchmark}$ = 0; all *p's* < .001). Thus, GenAI engages in significantly more stock picking when receiving a specific investment goal.

Total Expense Risk. Finally, we observe significantly higher total expense ratios when incorporating a social responsibility goal compared to our control condition and benchmark ($M_{Control}$= .1843%, $SD_{Control}$= .1967%; M_{Goal}= .3186%, SD_{Goal}= .1354%; $M_{Benchmark}$ = .07%; all *p's* < .001).

Figure 4. Incorporating a social responsibility goal in the prompt causes overleverage in utilities (Panel A) and decreases risk of low ESG rating (B) (Study 3).



Discussion

GenAI models often reveal biases implicitly present in the underlying training data [9,19–22]. Our findings suggest that these biases can result in generative investment advice that leads to elevated investment risks for private investors across five risk dimensions (geographical cluster risk, sector cluster risk, trend chasing risk, active investment allocation risk, and total expense risk) and that a range of debiasing interventions only partially mitigated these risks. We find that GenAI systems recommend a narrow range of geographical regions to invest, concentrate on a few narrow sectors, engage in more aggressive trend-following and stock-picking, and ultimately propose high-fee investment options—all recommendations that are incompatible with modern portfolio theory [15]. These findings illustrate the need for new risk frameworks in algorithmic finance and to better understand underlying data biases that can lead to a reinforcement of traditional (human) investment biases.

From a practical perspective, the findings of this research may serve as a critical guide for private investors who already started to employ generative AI systems to receive financial advice [4]. We hope that the current findings highlight the potential risks from a portfolio standpoint and promote a more critical assessment of the advice they receive. However, it is noteworthy to highlight that a critical assessment becomes more difficult when generative AI investment advice is communicated very convincingly (see Appendix B for details). These findings are sensitive because they suggest a potentially misleading sense of confidence and benevolence by AI-generated investment advice. This finding is important, in light of prior work showing that GenAI advice is often perceived as strikingly more positive even though the advice might carry inherent bias or when the advice is objectively worse compared to expert humans [23–28].

Methods

Experimental Paradigm. Our baseline experimental paradigm employs a large-scale experimental setting using a 3 (risk tendency: high / medium / low) x 3 (age: 15 / 30 / 50) x 3 (GenAI: ChatGPT / Copilot / Gemini) x 10 (number of queries per experimental condition) full factorial design. The first two experimental factors (risk tendency and age) provide our baseline experimental paradigm, representing private investors' self-stated risk-taking tendency and their objective risk exposure (e.g., older vs. younger investors should generally invest in less risky assets due to their limited ability to equalize negative returns over the investment horizon; [29,30]. The third and fourth experimental factors (type of GenAI and number of iterations) address concerns raised in earlier GenAI research, highlighting potential sampling issues when receiving and analyzing single (and sometimes less common) GenAI responses [31]. We therefore used the three predominant GenAIs available and performed automated API queries for each GenAI (ChatGPT, Copilot, Gemini) for a total of ten times to avoid potential selection issues when generating single prompts. We accessed ChatGPT 3.5 via OpenAI's API [32], Copilot via the EdgeGPT Python package [33], and Gemini via the Bard API Python package [34]. We focused on these large language models (LLMs) as they represent the predominant GenAI platforms in the marketplace. To ensure open science research practices [35], our code is documented and available on the Open Science Framework along with all datasets (https://osf.io/6mwk3/?view_only=e5a8b5133cad406c889280a90f0e948f). The specific baseline prompt template with our experimental variations across all studies was performed as follows: "*I'm a <15-year-old / 30-year-old / 50-year-old> looking to invest $10,000. I'm <not willing to take too many risks / willing to take some risks / willing to take a lot of risks> with my investment, and I'm hoping for some advice on what products I should consider investing in and how much I should allocate to each. Can you provide me with some recommendations? Please provide me with a table with the type of the investment, the name, the ticker symbol, and the amount I should invest.*". The next section details how each GenAI response was further processed and how each key measure of interest was computed from the GenAI's unstructured text response (see Appendix E for exemplary GenAI response).

Text Parsing & Data Augmentation. The table output received by each GenAI (see preceding section) was parsed and further augmented with additional financial data as follows: First, we converted the unstructured text table into a structured dataset by splitting the received key information into their core parts (type of investment such as "Stocks", name of the investment option such as "Apple Inc.", ticker symbol such as "AAPL", and amount in USD such as "$3,000"). We performed regular expressions to split each GenAI's tabular response into these four baseline categories. Second, we then further augmented the received investment advice with two key sources. Specifically, we augmented our baseline dataset by additional labeled data using Yahoo Finance (via the yfinance Python package [36]) and the Refinitiv Eikon database [37] to receive asset type, geolocation, sector, TER, as well as ETF and mutual fund holding data. For a comprehensive overview of the exclusion criteria and robustness checks, see Appendix F.

Each query is first augmented by additional data labels using Yahoo Finance (such as asset type) and then followed by augmenting the data further by adding performance metrics for each investment option such as the asset's TER (See Appendix G for a summary of our data acquisition and augmentation approach). To establish a risk-free baseline, we also collected FRED data [38] to retrieve risk-free market rate at the time of this study (i.e. 10-year US treasury bill).

Measurement. To assess the risk of the received recommendation, we developed a comprehensive GenAI investment risk model. We quantify investment risks along five key dimensions: Geographical cluster risk, sector cluster risk, trend chasing risk, active investment allocation risk, and total expense risk.

Table 1 provides a summary of each measure, the relevance from an investor perspective, and our empirical measurement approach for each type of risk assessment.

Benchmark. We contrast the received financial advice relative to one of the most commonly used financial benchmark indices 39. Specifically, we contrast all recommendations received by each GenAI relative to the Vanguard Total World Stock Index Fund (VT) ETF. This ETF represents a basket of securities that track the underlying index FTSE Global All Cap. Notably, the FTSE Global All Cap index serves as a strategic benchmark for the Norwegian Government Pension Fund Global, which holds 1.5% of the world's listed companies 40. In short, this benchmark represents a common, broad, and diversified investment portfolio. As with the data augmentation strategy, this data was queried using Yahoo Finance and the Refinitiv Eikon database.

References

1. Hildebrand, C. & Bergner, A. Conversational robo advisors as surrogates of trust: onboarding experience, firm perception, and consumer financial decision making. *J. Acad. Mark. Sci.* **49**, 659–676 (2021).
2. Capponi, A., Ólafsson, S. & Zariphopoulou, T. Personalized Robo-Advising: Enhancing Investment Through Client Interaction. *Manag. Sci.* **68**, 2485–2512 (2022).
3. Holzmeister, F., Holmén, M., Kirchler, M., Stefan, M. & Wengström, E. Delegation Decisions in Finance. *Manag. Sci.* **69**, 4828–4844 (2023).
4. McGowan, P., Harries, E., Manioti, L., Barnes, C. & Nunneley, S. *The Investor Index*. (2023).
5. Boukherouaa, E. B. & Shabsigh, G. *Generative Artificial Intelligence in Finance: Risk Considerations*. https://www.imf.org/en/Publications/fintech-notes/Issues/2023/08/18/Generative-Artificial-Intelligence-in-Finance-Risk-Considerations-537570 (2023).
6. Srivastava, A. *et al.* Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. Preprint at https://doi.org/10.48550/arXiv.2206.04615 (2023).
7. DeVries, T., Misra, I., Wang, C. & van der Maaten, L. Does Object Recognition Work for Everyone? Preprint at https://doi.org/10.48550/arXiv.1906.02659 (2019).
8. Big Data's Disparate Impact. *Calif. Law Rev.* (2016) doi:10.15779/Z38BG31.
9. Chowdhery, A. *et al.* PaLM: Scaling Language Modeling with Pathways. Preprint at https://doi.org/10.48550/arXiv.2204.02311 (2022).
10. Fang, L. & Peress, J. Media Coverage and the Cross-section of Stock Returns. *J. Finance* **64**, 2023–2052 (2009).
11. Statista. *Share of Value Added to the Gross Domestic Product of the United States in 2022, by Industry*. https://www.statista.com/statistics/248004/percentage-added-to-the-us-gdp-by-industry/ (2023).
12. Barber, B. M., Huang, X., Odean, T. & Schwarz, C. Attention-Induced Trading and Returns: Evidence from Robinhood Users. *J. Finance* **77**, 3141–3190 (2022).

13. Schaal, E. & Taschereau-Dumouchel, M. Herding through Booms and Busts. *J. Econ. Theory* **210**, (2023).

14. Gârleanu, N. & Pedersen, L. H. Active and Passive Investing: Understanding Samuelson's Dictum. *Rev. Asset Pricing Stud.* **12**, 389–446 (2022).

15. French, K. R. Presidential Address: The Cost of Active Investing. *J. Finance* **63**, 1537–1573 (2008).

16. Wermers, R. Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transactions Costs, and Expenses. *J. Finance* **55**, 1655–1695 (2000).

17. Facebook. facebook/bart-large-mnli · Hugging Face. https://huggingface.co/facebook/bart-large-mnli (2023).

18. Li, P., Castelo, N., Katona, Z. & Sarvary, M. Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis. *Mark. Sci.* (2024) doi:10.1287/mksc.2023.0454.

19. Brown, T. *et al.* Language Models are Few-Shot Learners. in *Advances in Neural Information Processing Systems* vol. 33 1877–1901 (Curran Associates, Inc., 2020).

20. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).

21. Nadeem, M., Bethke, A. & Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* 5356–5371 (Association for Computational Linguistics, Online, 2021). doi:10.18653/v1/2021.acl-long.416.

22. Achille, A., Kearns, M., Klingenberg, C. & Soatto, S. AI model disgorgement: Methods and choices. *Proc. Natl. Acad. Sci.* **121**, e2307304121 (2024).

23. Ayers, J. W. *et al.* Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern. Med.* **183**, 589–596 (2023).

24. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

25. Cadario, R., Longoni, C. & Morewedge, C. K. Understanding, explaining, and utilizing medical artificial intelligence. *Nat. Hum. Behav.* **5**, 1636–1642 (2021).

26. Celiktutan, B., Cadario, R. & Morewedge, C. K. People see more of their biases in algorithms. *Proc. Natl. Acad. Sci.* **121**, e2317602121 (2024).
27. Böhm, R., Jörling, M., Reiter, L. & Fuchs, C. People devalue generative AI's competence but not its advice in addressing societal and personal challenges. *Commun. Psychol.* **1**, 1–10 (2023).
28. Strachan, J. W. A. *et al.* Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* 1–11 (2024) doi:10.1038/s41562-024-01882-z.
29. Barberis, N. Investing for the Long Run when Returns Are Predictable. *J. Finance* **55**, 225–264 (2000).
30. Wachter, J. A. Portfolio and Consumption Decisions under Mean-Reverting Returns: An Exact Solution for Complete Markets. *J. Financ. Quant. Anal.* **37**, 63–91 (2002).
31. Chen, Y., Andiappan, M., Jenkin, T. & Ovchinnikov, A. A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.4380365 (2023).
32. OpenAI. OpenAI Developer Platform. (2023).
33. Cheong, A. EdgeGPT. (2023).
34. Park, M. Google Bard API. (2023).
35. Fišar, M., Greiner, B., Huber, C., Katok, E. & Ozkes, A. I. Reproducibility in Management Science. *Manag. Sci.* **70**, 1343–1356 (2024).
36. Aroussi, R. yfinance. (2023).
37. Refinitiv. Refinitiv Eikon. https://eikon.refinitiv.com/ (2024).
38. Federal Reserve. Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity, Quoted on an Investment Basis. *FRED, Federal Reserve Bank of St. Louis* https://fred.stlouisfed.org/series/DGS10 (2023).
39. Mork, K. A., Eap, H. M. & Haraldsen, M. E. Portfolio Choice for a Resource-Based Sovereign Wealth Fund: An Analysis of Cash Flows. *Int. J. Financ. Stud.* **8**, (2020).
40. Norges Bank. The fund. *Norges Bank Investment Management* https://www.nbim.no/en/ (2024).

Author Contributions

PW: Conceptualization, data collection, data analysis, writing – original draft; CH: Conceptualization, data analysis, supervision, writing – original draft, writing – review and editing; JH: supervision, writing – review and editing

Data Availability Statement

The datasets generated during and/or analyzed during the current study are available on the Open Science Framework, https://osf.io/6mwk3/?view_only=e5a8b5133cad406c889280a90f0e948f.

Competing Interests

The authors declare no competing interest.

Figure Legends

Figure 1.

Figure 2.

We find above benchmark geographical cluster risk (Panel A; N=269), sector cluster risk (Panel B; N=269), trend chasing risk (Panel C; N=270), and active investment risk (Panel D; N=270). The violin plots and boxplots represent the shape of the distribution of the respective characteristic by GenAI. The black dot represents the mean and the colored dots the value of individual portfolios. The dashed line represents the corresponding value of the benchmark.

Figure 3.

We find above benchmark geographical cluster risk (Panel A; N=180), sector cluster risk (Panel B; N=180), active investment risk (Panel D; N=180), and non-different trend chasing risk (Panel C; N=180) when using a broad debiasing intervention. The violin plots and boxplots represent the shape of the distribution of the respective characteristic by condition. The black dot represents the mean and the colored dots the value of individual portfolios. The dashed line represents the corresponding value of the benchmark.

Figure 4.

Panel A illustrates the portfolio weight by condition and country relative to the benchmark for the top five sectors (N=180). In panel B (N=180) the violin plots and boxplots represent the shape of the distribution of the respective characteristic by condition. The black dot represents the mean and the colored dots the value of individual portfolios. The dashed line represents the ESG score value of the benchmark.

Tables

Table 1. Overview of the risk measures.

Risk Measure	Definition	Consumer Relevance	Empirical Identification	Mathematical Formulation
Geographical cluster risk	The risk of overexposure to a specific geographic region.	May lead to amplified losses during region-specific economic downturns. Reduced diversification can increase volatility in the portfolio.	Mean proportion of money invested in the US vs. other countries to the total amount invested in investment assets of a portfolio	$CR_{Geo}^{US} = Mean\left(\frac{Money\ invested\ in\ US}{Money\ invested}\right)$
Sector cluster risk	The risk of over-investment within a particular sector or industry.	Limits the portfolio's exposure to potential gains from other sectors and increases sensitivity to sector-specific downturns.	Mean Sector Herfindahl-Hirschman index	$CR_{Sector} = Mean\left(\sum_{i=1}^{N}\left(\frac{Money\ invested\ in\ Sector_i}{Money\ invested}\right)^2\right)$
Trend chasing risk	The risk incurred by following recent market trends in investment decisions.	May result in buying high and selling low, leading to suboptimal returns due to entering. and exiting positions at non-ideal times.	Mean proportion of the amount of the top three assets by volume (when included in the top 20 most traded by volume three months before our data collection) to the total investment amount of a portfolio	$TCR = Mean\left(\frac{Money\ invested\ in\ top\ 3}{Money\ invested}\right)$
Active investment allocation risk	The risk of underperformanc e relative to a benchmark due to active management.	Higher transaction costs, potential for human error, and style drift can lead to underperformance compared to passive strategies.	Mean proportion of amount invested in actively managed assets (equities, bonds, cryptocurrencies, money market investments, seed investments, real estate investments were labelled as active due to their investment nature requiring active management) to amount invested in all assets of a portfolio	$AIAR = Mean\left(\frac{Money\ actively\ invested}{Money\ invested}\right)$
Total expense risk	The risk that high total expenses (such as management fees and operational costs) will diminish net investment returns.	Affects compounding potential of investments, potentially leading to significantly lower wealth accumulation over time. High costs are particularly detrimental in low-return environments.	Mean of the mean TER of ETFs and mutual funds of a portfolio	$TERR = Mean(TER)$