AI and Perception Biases in Investments: An Experimental Study[*]

Anastassia Fedyk
UC Berkeley

Ali Kakhbod
UC Berkeley

Peiyao Li
UC Berkeley

Ulrike Malmendier
UC Berkeley, NBER, and CEPR

Abstract

AI promises to accelerate and broaden access to automated investment advice. But can it capture the investment preferences and rationales of historically underrepresented investors? We ask 1,272 human survey respondents and 1,350 AI-generated agents to rate stocks, bonds, and cash. First, default AI-generated responses overrepresent the preferences of young high-income individuals. However, algorithmic bias disappears with demographically-seeded prompts. Second, AI-generated free-form responses correctly reflect human rationales: risk and return, financial knowledge, and past experiences. Third, AI can help identify where a lack of financial knowledge poses issues in human responses, as shown in textual analyses of transitivity violations.

Keywords: Investment preferences, Large language models, Behavioral biases, Experimental economics, Financial surveys, Generative AI.

JEL Classification: C1, G10, G11, G12.

[*] This Draft: August 2025. First Draft: March 2024. Anastassia Fedyk, Ali Kakhbod, and Peiyao Li are with the Haas School of Business at UC Berkeley. Ulrike Malmendier is with the Department of Economics and the Haas School of Business at UC Berkeley, NBER, and CEPR. E-mails: Fedyk (fedyk@berkeley.edu), Kakhbod (akakhbod@berkeley.edu), Li (ojhfklsjhl@berkeley.edu), Malmendier (ulrike@berkeley.edu). This paper supersedes *ChatGPT and Perception Biases in Investments: An Experimental Study*. We are grateful to Elena Asparouhova, Leland Bybee, James Choi, Francesco D'Acunto, Amir Kermani, Martin Lettau, Dmitry Livdan, Alejandro Lopez-Lira, Shumiao Ouyang, Alberto Rossi, Michael Reher, Gregor Schubert, Stas Sokolinski, and participants at BI Norwegian Business School, CUNY Baruch, NYU, Stanford University, UC Berkeley, University of Nevada-Reno, ASU Sonoran Winter Finance Conference, Carey Finance Conference, European Finance Association, NBER Behavioral Finance, and SITE Experimental Economics for very helpful comments.

1 Introduction

The financial services industry is undergoing rapid automation, with robo-advising and online trading platforms such as Robinhood assuming key roles in the investment landscape (Barber et al., 2022; D'Acunto, Prabhala and Rossi, 2019). Over the past decade, automated investment advice has grown more than tenfold, and declining account limits have increased access for less wealthy households (Reher and Sokolinski, 2024). At the same time, the economy (and specifically the financial sector) is amid a drastic transformation driven by recent AI technologies (Bybee, 2023; Cao et al., 2021; Erel et al., 2021; Fedyk et al., 2022; Lopez-Lira and Tang, 2023; Lyonnet and Stern, 2024). AI-driven systems, particularly large language models, have the potential to further expand access to low-cost automated investment products. However, a prerequisite for the eventual deployment of AI-powered financial advisors is the ability of AI systems to capture the diverse investment preferences of increasingly heterogeneous investors.

The promise of attracting and supporting more diverse and lower-income investor populations gives rise to a fundamental concern: Can AI systems adequately represent the preferences of populations that have been historically underrepresented among investors, even though AI agents are inherently backward-looking? Or will algorithmic bias hinder representation of these minority groups? This concern is particularly pertinent as, in the past, investor populations have featured a large imbalance in demographic characteristics such as gender (Barber and Odean, 2001). For example, low stock-market participation is an enduring concern among female, older, and lower-income individuals (Guiso, Sapienza and Zingales, 2008; Hong, Kubik and Stein, 2004; Van Rooij, Lusardi and Alessie, 2011). In lending, the introduction of machine learning has been found to disproportionately benefit white borrowers (Bartlett et al., 2022; Fuster et al., 2022).

In this paper, we evaluate the extent to which state-of-the-art generative AI models can replicate human investment preferences and the underlying rationales. We conduct a large-scale survey with 1,272 human participants and 1,350 generated AI agents responding to the same questions, with and without providing demographic information to the AI model. First, we find that, when asked to rate investments in stocks, bonds, and cash, default AI-generated responses overrepresent the preferences of young high-income individuals. However, when demographic attributes are incorporated into the prompts, AI captures the diversity observed in human responses, achieving a 70% correlation. Thus, algorithmic bias in predicting investment preferences appears to be avoidable with appropriately specified prompts.

Understanding investment preferences entails more than the correct ranking of asset classes; it also requires an alignment of the underlying rationales. In our second set of results, we compare the free-form explanations of investment choices provided by human respondents and AI agents. We find that AI-generated responses closely capture the stated rationales behind human responses. Specifically, they mirror how risk and return influence investment ratings and how the impact of financial knowledge and past experiences varies across demographics. These results suggest that AI systems have the potential to assess the roles of risk preferences, information, and distorted beliefs when internalizing the investment preferences of investors from different demographic groups.

Third, we utilize the evidence on non-transitive preference orderings to explore whether AI can help identify where investment choices are hampered by the lack of financial knowledge. When we elicit pairwise comparisons of investment options, AI exhibits transitive orderings 98.7% of the time, compared to only 84.4% for humans. Textual analysis of free-form explanations shows that violations of transitivity among human survey participants are associated with a lack of financial knowledge, expressed as indifference between at least two investment options. This analysis illustrates how AI systems can help identify where the lack of financial knowledge induces confusion or uncertainty about investment choices, which can potentially be remedied by providing information or financial literacy training.

Our human survey was administered in two waves, October 2023 and March 2024, and consisted of two sections: (i) categorical ratings of three investment options (stocks, bonds, and cash) on a scale from 1 (“very negative”) to 5 (“very positive”); and (ii) a comparison of each pair of investment options (stocks versus bonds, stocks versus cash, and bonds versus cash). For each question, the respondents were also asked to provide a free-form explanation. The human respondents are balanced across key demographics, with 49.8% identifying as male, a median age of 37, and a median personal income of $53,000.

We queried OpenAI’s GPT4 model with the exact same survey questions, describing the type of questions and permitted answers (rating versus free-form) in the prompt. We queried 150 simulated runs without providing demographic information in the prompt. In another 1,200 simulated runs (150 per demographic group), we seeded the prompts with gender (male or female), age (above or below the U.S. Census median of 39), and annual income (above or below the U.S. Census median of $54,000). In robustness analyses, we confirm that our results are robust to using different AI models, varying the level of randomness in the responses, and changing the prompt design.

We begin by comparing generative AI's rating patterns against the human ones *in the absence* of explicitly specifying the demographics in the AI model—that is, using the responses of the 150 default AI agents. To do so, we split human responses into eight demographic groups categorized by gender (men or women), age (above or below the median age in the U.S. Census), and income (above or below the median income in the U.S. Census). We model ratings for stocks, bonds, and cash for each of the eight human demographic groups using three-dimensional Gaussian distributions. Then, for each of the 150 simulated AI agents, we identify the demographic group whose rating distribution is most closely aligned with the responses of that agent. Default AI-generated responses match the profile of young high-income individuals in 83% of the cases, although this group comprises less than 30% of the population.

Given the link of algorithmic bias to the behavior of the representative investor, we test whether algorithmic bias is mitigated by providing demographic information to the model. We leverage the responses to the same questions from 1,200 AI agents with seeded demographics (150 per demographic group). We construct two 24-dimensional vectors—one from the human survey data and one from AI-generated responses—capturing average ratings for stocks, bonds, and cash from each of the eight demographic groups. The resulting Pearson (0.73) and Spearman (0.70) correlations indicate substantial correspondence between AI and human responses. Zooming in on rating differences across demographic groups, we confirm that men rate stocks higher than women, consistent with prior evidence (Agnew, Balduzzi and Sunden, 2003). Higher-income individuals rate stocks and bonds more favorably and cash less favorably than lower-income individuals (Hong, Kubik and Stein, 2004), and older participants assign higher ratings to cash than younger participants. AI-generated responses capture this heterogeneity, consistently across different AI models, prompts, and levels of randomness.

The demographic patterns in the AI data are also consistent with real-world asset allocations. We leverage the 2022 wave of the Survey of Consumer Finances (SCF) to validate our AI-generated investment preferences against real-world demographic patterns in asset allocation. The SCF data confirm that AI accurately captures key allocation trends: women maintain higher cash reserves, older investors favor bonds, and high-income individuals prefer stocks while holding less cash.

Our results suggest that algorithmic bias in the default AI responses is driven by the model "assuming" a particular demographic identity. We develop a novel data injection experiment to test this hypothesis. We provide the model with a few (5, 10, or 20) sample responses from the demographically-seeded data but with the ratings flipped ("fake data"), so that the demographic associations are reversed. This experiment succeeds in flipping the output of the model, which begins to copy the young high-income group in the "fake data" and favor low-growth investments. Thus, algorithmic bias in default AI agents stems from the tendency to adopt a young high-income identity, rather than an intrinsic proclivity for high-growth assets independently of demographic cues.

In the second part of the paper, we examine the factors driving investment ratings by drawing insights from the free-form explanations accompanying each rating, both among humans and among AI agents. We introduce an unsupervised text-based approach to investigate the most common explanations for a given investment preference. We use a combination of (i) generative AI to identify the main themes in free-form responses and (ii) a textual embedding technique based on the Semantic Axis (SemAxis) framework in the natural language processing literature (An, Kwak and Ahn, 2018) to quantify the loading of each open-ended response on each theme.

The two most prevalent themes in both human and AI-generated responses are risk and return. We document that the mappings between the risk and return perceptions and the ratings of stocks, bonds, and cash are very similar between human survey responses and AI-generated responses. A one-standard-deviation increase in the return dimension corresponds to a 0.79-standard-deviation increase in the human ratings and a 0.78-standard-deviation increase in AI-generated ratings. A one-standard-deviation increase in risk perception translates into a 0.56-standard-deviation lower rating in both human and AI-generated data. These findings suggest the potential for generative AI tools to correctly capture the rationales behind expressed preferences.

Beyond risk and return, our textual analysis approach identifies the two main themes behind differing attitudes towards the stock market as (i) knowledge and understanding of the stock market (Van Rooij, Lusardi and Alessie, 2011), and (ii) personal experiences with investing in the stock market and the resulting emotions (Malmendier and Nagel, 2011, 2016). We show that generative AI matches the demographic differences in these factors. In terms of knowledge, men and high-income individuals are more likely to express familiarity with the stock market. In terms of past experiences, young individuals, men, and high-income earners discuss more positive experiences and emotions. Moreover, when we examine the *associations* between the four key themes (risk, return, knowledge, and experiences) in the free-form explanations of stock ratings, we find very similar patterns in human and AI-generated responses. Knowledge and past experiences are positively associated with each other, positively associated with return perceptions, and negatively associated with risk perceptions.

Overall, generative AI seeded with demographic information captures the underlying "factor structure" of investment preferences. Not only does the model accurately predict first-order moments such as differences in investment ratings between men and women, but it also replicates the themes that drive these differences, and how these themes relate to each other. Instead of generated responses capturing one factor at a time (e.g., some positive ratings of stocks reflecting high return expectations, and others reflecting positive past experiences), AI reproduces the *associations* between these factors within-response.

The ability of generative AI models to capture heterogeneous investment preferences *and* the sources of this heterogeneity suggests that AI has the potential to serve an increasingly broad investor base. With properly seeded prompts, generative AI can correctly account for heterogeneous preferences (e.g., different risk aversion parameters) while potentially screening out biases (e.g., beliefs that over-infer from past experiences).

In the final part of the paper, we focus on one such bias, prevalent among human respondents but not in the AI-generated data: violations of transitivity. In the direct-comparison responses, where participants compare pairs of investment options, 98.7% of AI-generated preference orderings are transitive, compared to 84.4% of human preference orderings. This difference highlights AI's ability to maintain logical consistency in investment preferences, a critical feature for financial decision-making tools. In the human response sample, we find that violations of transitivity are more common among women, and most of them occur in individuals who express indifference between at least two investment options. Our textual analysis of the free-form explanations accompanying relative comparison questions reveals that a key theme is the lack of financial knowledge. Survey participants express their lack of knowledge as indifference that breaks transitivity, as we confirm through mediation analysis. This suggests that transitivity of preference orderings is one area in which AI can help debias human preferences.[1] More generally, AI can help uncover reasons for biases from free-form explanations.

Related literature. Our results contribute to the rapidly growing literature on the effects of financial technology on the investing landscape. The availability of low-commission

[1] Responses generated by AI tend to preserve other biases, including in expectation formation (Bybee, 2023), making the debiasing result in transitivity especially intriguing.

online trading platforms, such as Robinhood, has expanded retail investor participation in the stock market, while the rise of robo-advising has increased the prevalence of automated investing (Barber et al., 2022; D'Acunto, Prabhala and Rossi, 2019; Welch, 2022). These trends underscore the importance of understanding how technological advances will affect an investment landscape comprised of increasingly diverse participants. To date, robo-advising has led to greater portfolio diversification (D'Acunto, Prabhala and Rossi, 2019) and mitigation of certain biases (D'Acunto, Ghosh and Rossi, 2022). However, the welfare gains have been unevenly distributed across demographics such as investor age (Reher and Sokolinski, 2024; Rossi and Utkus, 2021). Coupled with the recent evidence of algorithmic bias in domains ranging from medicine (Kadambi, 2021) to credit supply and loan interest rates (Bartlett et al., 2022; Fuster et al., 2022), this raises concerns that new advances in artificial intelligence—such as the use of large language models trained on text from the Internet—may disproportionately reflect the investment preferences of specific demographic groups (e.g., young men).

Our results directly address those concerns, showing that (i) the default model is biased, but (ii) the bias disappears when the model is seeded with demographic information, and (iii) the underlying structure of the knowledge base—including factors that drive heterogeneous preferences—resembles true human reasoning. This has important implications for automating financial advice. A successful AI-driven advisory system must both understand individual investors and guide them effectively. Our research confirms that AI excels in the first aspect by accurately capturing investor heterogeneity and preference structure. Future research can explore AI's capacity to not only understand but also improve decision-making by optimizing investment allocations and striking a balance between personalization and de-biasing. Our analysis of transitivity takes a first step in this direction.

Our findings also contribute to the emerging literature on the effects of AI on various sectors of the economy, with our work speaking to the financial services sector.[2] A growing line of work shows that AI can be effectively incorporated into financial and business practices if proper considerations are taken into account (Bybee, 2023; Hansen and Kazinnik, 2023; Kakhbod et al., 2024; Lopez-Lira and Tang, 2023).[3] We build on this

[2] The effects of generative AI on labor and firms have been studied by Bertomeu et al. (2023); Brynjolfsson, Li and Raymond (2023); Eisfeldt et al. (2023); Eloundou et al. (2023); Noy and Zhang (2023), among others.

[3] Outside of finance, Horton (2023) shows that generative AI behaves similarly to human experimental participants, with analogous responsiveness to endowments.

work by examining generative AI's ability to capture individual-level heterogeneity in investment preferences and the underlying reasons driving this heterogeneity. Our results offer a promising starting point for leveraging generative AI to understand the drivers of heterogeneity in human preferences, especially when direct survey evidence may be difficult to obtain. AI's ability to reflect the underlying structure of explanatory themes can make it a powerful tool for hypothesis formation and exploring underlying mechanisms.

Finally, we offer a methodological contribution by introducing a two-step, flexible, and cost-effective approach to extract meaningful insights from survey data—including free-form textual explanations—in an unsupervised manner. Traditional survey methods require researchers to form ex-ante predictions and design survey questions that ask participants to rate each listed factor, which can constrain the range of observed behaviors. To address this limitation, our first step begins with open-ended text responses and applies a generative language model to parse these explanations and extract underlying themes directly from the data. This is especially valuable in financial applications, where the factor structure driving behavioral patterns is complex. In the second step, we quantify the extent to which each response maps to each theme, enabling the application of standard econometric techniques to analyze the extracted factors. Together, these steps provide a framework for leveraging rich, unstructured survey data to reveal insights that traditional methods might overlook (Haaland et al., 2024). Our methodological contribution builds on the agenda of using machine learning and AI tools to study unstructured financial data (Fedyk and Hodson, 2023; Loughran and McDonald, 2014; Tetlock, 2007) and glean insight from open-ended survey questions (Stantcheva, 2020).

The remainder of the paper proceeds as follows. We describe the data collection procedures in Section 2. We compare the ratings from the human survey and the generative AI model in Section 3 and then analyze the explanations in Section 4. Section 5 examines the transitivity of preference orderings. Section 6 concludes.

2 Data Collection: Human and Generative AI Surveys

We describe the methodology for collecting human investment preferences across demographics and analogous AI-generated responses.

2.1 Human survey

Human survey responses come from a representative sample of 1,272 individuals recruited through the Prolific platform. The survey was conducted in October 2023 and March 2024.[4] 1,264 individuals completed the entirety of the survey, including the demographic questionnaire at the end. We specified to the Prolific survey platform that all respondents had to be based in the U.S., and that the sample should be balanced on gender. The resulting sample was representative of the U.S. population on the demographic information we collected: gender, age, and income. 49.8% of the respondents reported being male, 47.5% identified as female, 2.3% chose "non-binary / third gender," and 0.5% declined to say. The median age of the respondents is 37, and the average is 39.7, comparable to the median age of the US population, which the US Census reports as 38.9 for 2022. The median income is $53,000 (with an average of $68,876), very close to the $54,339 median earnings for full-time year-round civilian employees in the US Census for 2021. Compensation for participating in the survey averaged $15.45/hour.[5]

The survey asked the respondents to evaluate three investment options—stocks, bonds, and cash—separately and relative to each other. The first three questions were single-rating questions, asking the respondents to rate each investment option on a scale of ″very negative″ (encoded as 1), ″somewhat negative″ (encoded as 2), ″neutral″ (encoded as 3), ″somewhat positive″ (encoded as 4), and ″very positive″ (encoded as 5). The three single-rating questions were presented in random order to avoid anchoring effects. Each of the three single-rating questions was followed by a free-form text entry question asking why the respondent chose that rating, which required responses with a minimum length of 20 characters. The purpose of these questions is to assess the underlying factors driving participants' ratings. Figure 1(a) in the Online Appendix displays an example single-rating question screen.

Next, the respondents faced three direct-comparison questions: whether they prefer stocks or bonds, stocks or cash, and bonds or cash. Each direct-comparison question presented three choices—corresponding to the two assets being compared and an option for indifference—and was accompanied by a free-form text entry asking "why?", with a minimum response length of 20 characters. The order of the direct-comparison questions

[4] An identical survey was administered on both dates, with an initial sample of 469 individuals in October 2023, and an additional 803 participants in March 2024.

[5] The survey payment was $1 and participants took, on average, approximately four minutes to complete the survey.

was randomized across participants. Furthermore, the order of the options *within* each question was also randomized across participants (e.g., half of the participants were asked whether they prefer ”stocks or bonds,” and the other half were asked whether they prefer ”bonds or stocks”) to avoid biasing the participants towards any options on aggregate. Figure 1(b) in the Online Appendix shows an example direct-comparison question. The exact instructions used in the survey are included in Online Appendix B.

We perform the following cleaning procedures on the survey data before commencing the analysis. First, we remove outliers in age and income using an inter-quartile range (IQR)-based rule. Specifically, a data point is considered an outlier if the participant's age or income is at least 1.5 IQRs above or below the median.[6] Second, we exclude human responses where the participant declined to disclose their gender or identified as non-binary because we lack sufficient data in the non-binary category to perform meaningful inference (these responses represent less than 3% of the sample). After these cleaning procedures, we retain a sample of 1,074 individuals with an average (median) age of 38 (36) and an average (median) income of $\$53,000$ ($\$50,000$), rounded to the nearest $1,000.

Since our paper compares human and AI-generated responses, we verify that our survey participants do not use AI tools to compose their free-form explanations.[7] First, we note that the length of the average human response is 12 words (1 sentence), and the overall survey is quite short, taking only four minutes to complete. Therefore, human participants have minimal incentive to consult AI tools for this task. Second, we use GPTZero, a state-of-the-art AI writing detector, to directly test whether each human response is likely to have been generated or modified by AI. For all questions (ratings of stocks, bonds, and cash, and pairwise comparisons of these asset classes), between 96% and 99% of the responses are identified as “human only” (the three potential labels are ”human only”, ”mixed”, and ”AI only”). This confirms that our sample of survey responses comes from real people and does not reflect the use of AI tools.

[6] We adopt the Tukey's outlier fence approach, which uses a 1.5 IQR rule by default (Tukey et al., 1977). We use the median as a reference point for both up and down, resulting in an upper limit of 67 for human age (the lower limit is 7, but all recruited survey participants are already 18 and over). Empirically, the outlier removal rule does not remove any human data points in our sample based on income.

[7] If survey participants used AI tools to write their explanations, this would introduce upward bias in the similarity between their responses and our sample of AI-generated responses.

2.2 AI data collection

We simulate survey data collection using AI to conduct a cleanly identified comparison between human and AI-generated investment preferences. As in the human survey, we elicit responses to three types of questions:

1. How do you rank each investment option (stocks, bonds, cash)?
2. What is your preference ordering among the three options (pairwise comparisons)?
3. What is your rationale (free-form explanation) for each rating and relative comparison?

For the first type of question, we give each simulated AI agent five choices, mirroring the human survey: very positive, somewhat positive, neutral, somewhat negative, and very negative. We again convert these ratings to numerical values from 1 (very negative) to 5 (very positive). For the second type of question (comparisons), we offer three choices: option 1, option 2, and indifferent. For example, when comparing stocks and bonds, the options are preferring "Stocks," preferring "Bonds," and "Indifferent." For the third type of question, we ask for a short explanation of each rating and comparison response using 5 to 10 words. Additionally, we ask each AI agent to report the gender, age, and income of its imagined identity. Online Appendix C presents a sample prompt.[8]

We perform the AI survey in two ways: (i) without specifying the demographics ex ante and (ii) seeding the AI model with different demographic characteristics. For the default generated responses not seeded with demographics, we query the model 150 times with the prompt provided in Online Appendix C. This sample of default responses gives us a benchmark of how generative AI behaves when it is *not* instructed to take on any particular identity. For the sample of AI responses seeded with demographic characteristics, we query the model 1,200 times (150 per demographic bin). In order to mimic the representative sample of participants in our human survey, we specify the median age and income according to the U.S. Census data discussed in Section 2.1 and query AI responses for imagined male or female agents above or below the median in age and income. AI can follow these instructions 100% of the time. Therefore, the male-to-female ratio is 50–50,

[8] In order to avoid introducing any bias in the AI-generated responses, we (i) execute each query separately and (ii) randomize the order of questions in each query, analogously to the randomization in the human survey. We set the default temperature hyperparameter to 0.8 to generate a random sample of responses with enough variation to match human survey participants. In robustness analyses, we confirm that the results are not sensitive to the exact choice of the temperature hyperparameter.

the average (median) age is 37 (39) years old, and the average (median) personal income is $\$56,000$ ($\$55,000$), rounded to the nearest $1,000. For comparability, we perform the same outlier removal procedure as we applied to the human survey responses. After processing, we retain responses from 1,042 AI agents with an average (median) age of 36 (36) years old and an average (median) income of $\$53,000$ ($\$53,000$), rounded to the nearest $1,000.

3 Human versus AI investment ratings

In this section, we study the promises and limitations of using AI-based approaches to replicate heterogeneous investment preferences across demographic groups. We first provide descriptive statistics of average ratings of stocks, bonds, and cash across different demographic groups of human survey participants. We then present the default ratings generated by AI in the absence of demographic information. Finally, we examine how well AI can match the patterns in human investment ratings when demographic information is seeded in the prompts.

3.1 Heterogeneity in the human data

We consider eight demographic groups, defined by three binary indicators: age, gender, and income. Gender is split into male and female.[9] Age and annual income are sliced at the median based on the U.S. Census: 38.9 years old and $54,339, respectively. As shown in Table 1, our sample of human participants is not concentrated in any particular group. The largest subgroup (young low-income men) accounts for 17% of the sample, while the smallest subgroups (old high-income women, old low-income men, and old high-income men) account for 10% each.

Table 1 also shows how investment preferences for stocks, bonds, and cash vary across demographic groups. For example, higher-income individuals tend to rate stocks and bonds more favorably than their lower-income counterparts. Older individuals assign higher ratings to cash than younger individuals, and women rate stocks lower than men.

[9] To preserve statistical power, we exclude the small number of observations (2.8%) with other gender options, see Section 2.1.

Gender	Age	Income	Stock	Bond	Cash	N
Female	Old	Low-income	3.52	3.56	3.28	130
			(0.08)	(0.07)	(0.10)	
Female	Old	High-income	3.83	3.70	3.28	103
			(0.10)	(0.08)	(0.12)	
Female	Young	Low-income	3.55	3.37	3.22	172
			(0.07)	(0.06)	(0.07)	
Female	Young	High-income	3.75	3.52	2.98	125
			(0.08)	(0.06)	(0.09)	
Male	Old	Low-income	3.75	3.57	3.29	106
			(0.10)	(0.09)	(0.10)	
Male	Old	High-income	3.95	3.77	3.13	110
			(0.10)	(0.08)	(0.11)	
Male	Young	Low-income	3.81	3.61	3.03	182
			(0.08)	(0.07)	(0.09)	
Male	Young	High-income	4.20	3.67	2.87	146
			(0.07)	(0.08)	(0.10)	

Table 1: This table displays the numbers of human survey participants across demographic groups based on gender, age, and income, the average ratings of stocks, bonds, and cash by each group, and the corresponding standard errors. Age and annual income are divided according to the U.S. Census: above and below 38.9 years old and $54,339, respectively.

3.2 Algorithmic bias in default AI agents

We compare the investment ratings of human survey respondents against those generated by default AI agents, based on 150 AI simulations without specified demographic attributes. On average, both groups assign the highest ratings to stocks, followed by bonds and then cash. Human respondents rate stocks at 3.8, bonds at 3.6, and cash at 3.1. Default AI agents exhibit a stronger preference for high-growth assets, with stocks and bonds rated at 4.6 and 4.0, respectively, while rating cash lower at 2.9.

To assess how the default AI ratings align with human survey responses, we construct three-dimensional Gaussian distributions based on the ratings of each human demographic group, which serve as benchmarks for comparison. We then evaluate each simulated AI agent by determining the most likely human distribution from which its ratings could have been drawn. The results indicate that a meager 0.667% of AI agents align most closely with the young low-income female group, while an identical 0.667%

align with the old high-income female group. A substantially larger share, 28.0%, corresponds to the young high-income female group. 7.333% of AI agents align with the old low-income male category, and 8.667% correspond to the old high-income male category. The largest share, 54.667%, is most closely matched with the young high-income male group. Consistent with these results, AI's self-reported demographics are also mostly young and high-income.[10]

Overall, we find that default AI agents' rating patterns overrepresent the young high-income segment of human participants, with a slight gender imbalance in favor of males, underscoring the potential for systematic bias when demographic attributes are not explicitly specified.

3.3 Source of algorithmic bias

There are two potential sources for the algorithmic bias whereby AI acts as the young high-income group and prefers high-growth assets (stocks and bonds over cash): an inherent preference for high-growth assets in the generative AI model, or a preference for mimicking the behavior of a certain demographic group (young high-income individuals), which happens to choose high-growth assets. In this section, we provide suggestive evidence for the latter interpretation.

We use a novel experimental intervention, where we inject a set of artificial training data ("fake data") into the AI model and then redo our main analysis. The artificial training data are formatted as question-answer pairs such as

> *Question: How would you rate investing in stocks (very positive, somewhat positive, neutral, somewhat negative, and very negative)? Also tell me your age and income.*
>
> *Answer: Age: XX; Income: YY; Rating: ZZ*

where XX and YY are the true observed self-reported age and income from a random subset of the AI-generated data with demographic groups specified in the prompt (described in more detail in Section 3.4 below), and ZZ is the *opposite* of the corresponding

[10] Our original prompt (see Online Appendix C.1) states "Imagine you are an online survey participant," which may bias the AI responses toward younger, digitally active demographics. To address this, we implement two robustness checks. In the first, we vary prompt to state "Imagine you are participating in a survey," removing any reference to the online medium. In the second, we vary the prompt to specify that the survey can be completed "online or by mail," balancing digital access with a traditional mailing option. In both variations, the AI model still overrepresents the investment ratings of young, high-income individuals, accounting for 79% of responses in the first variation and 83% in the second variation.

rating. For example, if a young high-income individual rated stocks as “very positive,” we would preserve the demographics but change the rating to “very negative.”

We inject these “fake data” into the AI model using in-prompt learning, where we randomly include K examples (question-answer pairs) in the prompt before posing the question.[11] We vary the intensity of this data injection by repeating the analysis for $K = 5, 10, 20$. An example prompt with five pairs is as follows:

> *Consider the following example question-answer pairs as additional data you observed in your training set: [Example 1]*
>
> *[Example 2]*
>
> $\vdots$
>
> *[Example 5]*
>
> *[remainder of the prompt as in Section 3.2.]*

This procedure artificially creates an opposing association between demographic characteristics and investment ratings: in the fake data, young high-income individuals rate stocks and bonds lower, while old low-income individuals rate them higher. The injection exercise allows us to observe model responses when the association is flipped. We test whether the model retains the preference for high-growth assets even if it means acting akin to the old low-income group, or whether it instead mimics the young high-income group even if it means favoring low-growth assets. In the former case, the additional data would not change the tilt towards high-growth assets (stocks and bonds) in the AI-generated responses, and the proportion of responses mapping to different human distributions would remain stable. In the latter case, AI-generated responses would begin to favor cash over stocks and bonds. Mapped to the human distributions, we would then observe that a larger proportion of AI agents align with the rating patterns of the old low-income group, in contrast to what we observed in Section 3.2.

Figure 1 presents the results of the corresponding mapping exercise. We follow the same procedure as in Section 3.2 to identify, for each simulated AI agent, the most similar group of human responses. In the two left-hand-side panels, we examine the portion of AI agents that map to the old low-income human groups (female in the top panel and male in the bottom panel), for different values of K. $K = 0$ corresponds to the baseline from Section 3.2: AI agents almost never resemble the preferences of old low-income

[11] The details of in-prompt learning can be found in Appendix D.2.

individuals. As K increases, we inject progressively more "fake data," and the proportion of AI agents mapping to the old low-income group rises. With $K = 20$ fake datapoints, the total (male + female) share of responses mapping to this group exceeds 50%. In contrast, the two right-hand-side panels of Figure 1 show that the share of responses mapping to the young high-income human preferences becomes progressively smaller as we provide additional fake data to the AI model.

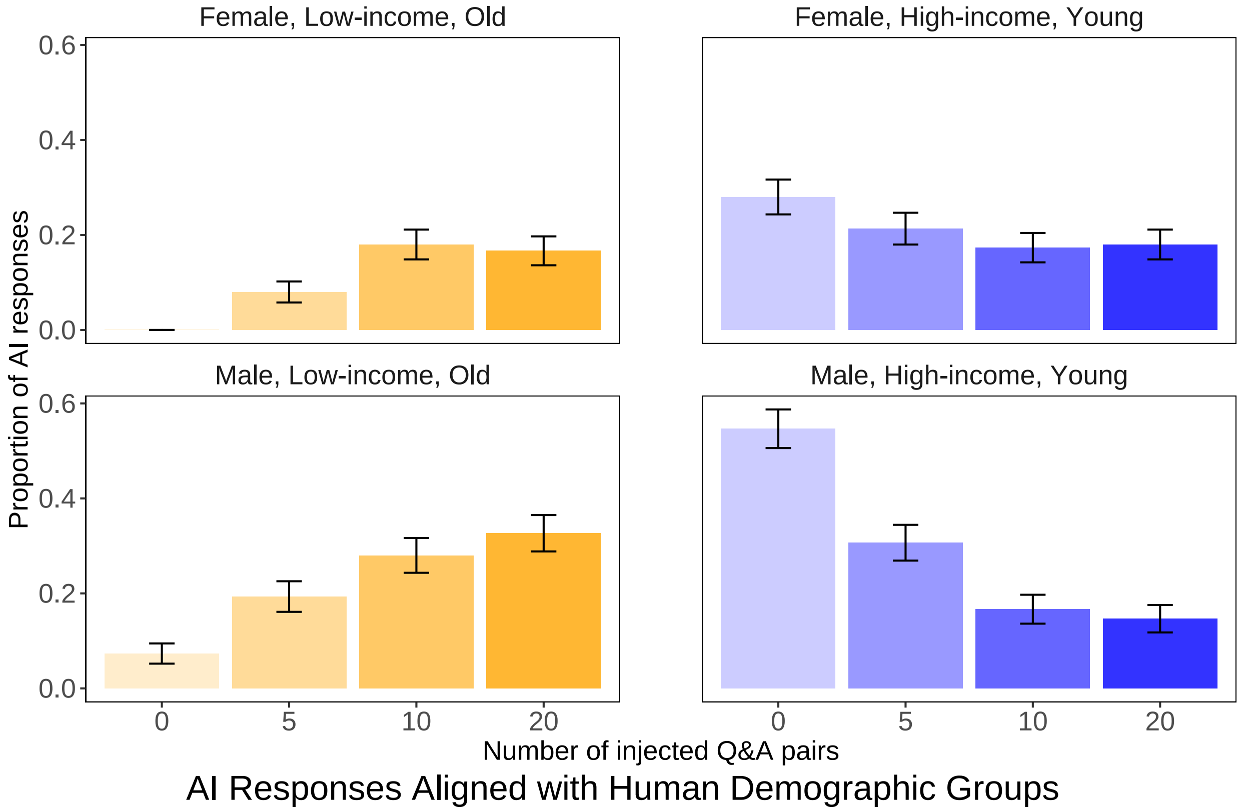


Figure 1: This figure shows the proportion of default AI agents' responses that best fit a given human demographic group, indicated in the title of the corresponding quadrant. The x-axis represents the number of additional question-answer pairs in the data injection experiment, where zero is the baseline without any injection. The two subplots on the left show the minority groups in the original AI data—old low-income groups—and the two subplots on the right show the majority groups— young high-income. This figure demonstrates an increasing (decreasing) representation of the minority (majority) groups when we inject samples that are engineered to resemble the opposite of the original default AI rating patterns.

This suggests that the default AI model follows the archetype of the young high-income group, and when additional training data (falsely) tell the model that this subset of people tends to rate stocks lower and cash higher (resembling old low-income individuals in the true human survey data), the AI model is more likely to mimic the same pattern, reflecting the updated (fake) preferences of young high-income individuals. This suggests that the bias in the default model stems from mimicking a specific demographic group (young high-income) rather than a direct preference for high-growth assets.

The injection exercise confirms that AI-based approaches to capturing investment preferences can overrepresent the empirically dominant group among investors, raising the concern of these systems' ability to serve traditionally underrepresented groups.

3.4 Bias mitigation with demographically seeded prompts

We now examine whether we can mitigate the bias by simply instructing AI agents to take on different pre-specified demographic identities. As detailed in Online Appendix C, we modify the prompt by including the following instructions at the beginning:

> *Imagine you are a [gender] online survey participant who is [age] old with an annual income [income].*

where gender is randomly assigned as male or female, age is either "below 39 years old and above 18 years old" or "at least 39 years old," and income is specified as either "above 54 thousand" or "at or below 54 thousand." All categories are assigned with equal probability. To ensure comparability with human survey data, we generate 150 responses from each of the eight demographic profiles, yielding a total of 1,200 AI-generated responses. Table A1 in the Online Appendix presents the summary statistics, which align with those in Table 1, with higher-income individuals favoring stocks and bonds more than lower-income individuals, while older individuals rate cash higher than younger individuals.

To assess the similarity between human and AI-generated ratings, we start by computing 24 average numerical ratings from each set of responses (human survey and AI): one average rating for each of the three investment options, separately for each of the eight demographic groups. We compute the correlations between the human and AI-generated ratings across these 24 groups. Table 2 shows the corresponding Pearson, Spearman, and Kendall correlations in the top row, which are very high at 0.73, 0.70, and 0.57, respectively. We compute the statistical significance of these correlation coefficients using a bootstrap of 10,000 samples of the same size as the original data drawn randomly with replacement from the human survey and AI-generated data. The bootstrapped standard errors are reported in parentheses in Table 2. All three correlation coefficients are highly statistically significant at the 1% level.

We also conduct the analysis separately within each asset class, focusing on the eight average ratings (across demographic groups) for stocks, bonds, and cash. These correlations are also very high. For stocks, the differential ratings across demographic groups from human and AI-generated responses show a Pearson correlation of 0.78, a Spearman correlation of 0.81, and a Kendall correlation of 0.71, all significant at the 1% level. Similarly, the ratings for cash show a Pearson correlation of 0.77, Spearman correlation of 0.64, and Kendall correlation of 0.64, all significant at the 1% level. The correlations for bonds are lower but still statistically significant, with a Pearson correlation of 0.58 (significant at the 1% level), Spearman correlation of 0.45 (significant at the 5% level), and Kendall correlation of 0.27 (significant at the 10% level).

	Pearson	Spearman	Kendall
All	0.728***	0.695***	0.565***
	(0.042)	(0.053)	(0.046)
Stocks	0.783***	0.810***	0.714***
	(0.099)	(0.125)	(0.132)
Bonds	0.581***	0.452**	0.286*
	(0.178)	(0.210)	(0.170)
Cash	0.768***	0.786***	0.643***
	(0.152)	(0.167)	(0.161)

Table 2: This table reports the Pearson, Spearman, and Kendall correlations between human and AI-generated responses. These correlations are computed based on the average ratings from each demographic group for the three investment options (pooled and separately). Bootstrapped standard errors are reported in parentheses.

To further examine the demographic patterns reflected in human and AI-generated data, we estimate regressions where the dependent variable is the rating (separately for stocks, bonds, and cash), and the independent variables are the demographic characteristics: gender, age, and income. We estimate the regression for human and AI-generated data separately to assess the extent to which the two sets of responses align in their demographic heterogeneity. We focus on coefficients that are statistically significant in both human and AI-generated responses and compare the direction of the associations.

Table 3 summarizes the results: five of the six significant coefficients go in the *same* direction in both human and AI-generated data.Specifically, older respondents tend to rate cash higher than younger ones, women tend to rate stocks lower than men, high-income individuals rate both stocks and bonds higher than low-income individuals, and high-income individuals rate cash less favorably than low-income individuals. The only case where AI responses do not correctly capture human investment preferences is the relationship between gender and bond ratings: the AI model expects women to rate bonds higher than men, whereas male human survey participants rate bonds higher than women. This one-off difference may be attributed to AI reflecting the standard investment allocation tradeoff between stocks and bonds (Agnew, Balduzzi and Sunden, 2003), rather than independently assessing each investment option.

	Human direction	AI direction	Agreement
old: cash	+	+	✓
female: stocks	−	−	✓
female: bonds	−	+	×
high-income: stocks	+	+	✓
high-income: bonds	+	+	✓
high-income: cash	−	−	✓

Table 3: This table displays the relationship between ratings of the investment options (stocks, bonds, and cash) and demographic characteristics (age, gender, and income). We present the six associations that are significant in both human and AI-generated responses. A plus (minus) sign in the second and third columns means that the corresponding demographic is positively (negatively) associated with the rated asset. For example, the first row shows that older individuals in both human and AI-generated data rate cash higher than their younger counterparts.

Overall, Tables 2 and 3 show that AI responses are qualitatively similar to human surveys, correctly reflecting most demographic differences in investment ratings. This confirms that providing specific demographic information mitigates the bias observed in the default AI model. In Appendix D, we show that this result is robust to using a different AI model from another provider (Anthropic), varying the prompt structure, and setting different levels of randomness.

3.5 External validity: Asset allocation

In this subsection, we examine the external validity and broader applicability of AI-based investment analysis by leveraging real household-level asset allocation data from the Survey of Consumer Finances (SCF) (Reher and Sokolinski, 2024).

The SCF is a comprehensive triennial survey conducted by the Federal Reserve Board in cooperation with the U.S. Department of the Treasury. The survey collects detailed information on the financial characteristics of U.S. households, including income, net worth, balance sheet components, pension plans, and other financial behaviors. We observe the age, gender, and income of each respondent. The SCF data are at the household level, which can include multiple income earners (e.g., married couples). We use the most recent survey wave from 2022. In our main specification, we restrict the sample to single-person households, resulting in a sample of 4,484 observations. However, the results are robust to including all households, marking demographics by the head of household, and using fixed effects for the number of people in each household. For the asset allocation, we observe the amount of money in savings and checking accounts (cash), the amount invested in publicly traded stocks, and the amount invested in fixed income (bonds). We compute the fraction allocated to each of these three asset classes as a share of the total.

We generate analogous asset allocations from AI using the prompt presented in Online Appendix C.3. For this elicitation, we used seeded prompts, specifying the age, gender, and income of a participant and asking for an allocation across the three asset classes adding up to 100%. Analogous to the SCF data, we calculate asset allocations across the three asset classes by age, gender, and income. The AI model allocates higher investment shares to stocks and bonds, compared to real human allocations in the SCF, which tend to hold cash. This is consistent with the low stock market participation puzzle (Guiso, Sapienza and Zingales, 2008) and suggests potential benefits to AI-guided investment advice. However, given our goal is to examine the ability of AI models to replicate heterogeneity in human preferences, we focus not on the levels but on the deviations in asset allocations across demographic groups in the AI-generated and SCF data.

In Table 4, we show the differences in asset allocations across demographic groups, analogous to the demographic differences in preferences reported in Table 3. We observe that older individuals invest more in stocks and bonds than younger individuals, women keep a larger proportion of their assets in cash than men, and high-income earners invest more in stocks and keep less cash than low-income earners. The responses generated by AI show the same patterns of demographic heterogeneity as the SCF data. These results confirm that large language models are able to match not only self-reported preferences but also how those heterogeneous preferences translate into real-world asset allocations.

In summary, we find a close alignement in the demographic differences in allocations between the actual human data (SCF) and AI-generated responses. This implies the potential of large language models for querying heterogeneous investment preferences in applications such as robo-advising, which we discuss further in Section 6.

	Human direction	AI direction	Agreement
old: bonds	+	+	✓
female: cash	+	+	✓
high-income: stocks	+	+	✓
high-income: cash	−	−	✓

Table 4: This table displays the associations between portfolio allocations to the three asset classes and demographic characteristics (age, gender, and income). We show the four associations that are significant in both real allocations (the Survey of Consumer Finances) and AI-generated data. A plus (minus) sign in the second and third columns means that the corresponding demographic is positively (negatively) associated with the allocation to the corresponding asset class.

4 Understanding the rationales of investors

In Section 3, we have shown that default large language models suffer from algorithmic bias, overrepresenting young high-income individuals. With demographic information in the prompt, however, the bias disappears, and AI-generated responses align closely with human survey preferences and asset allocations.

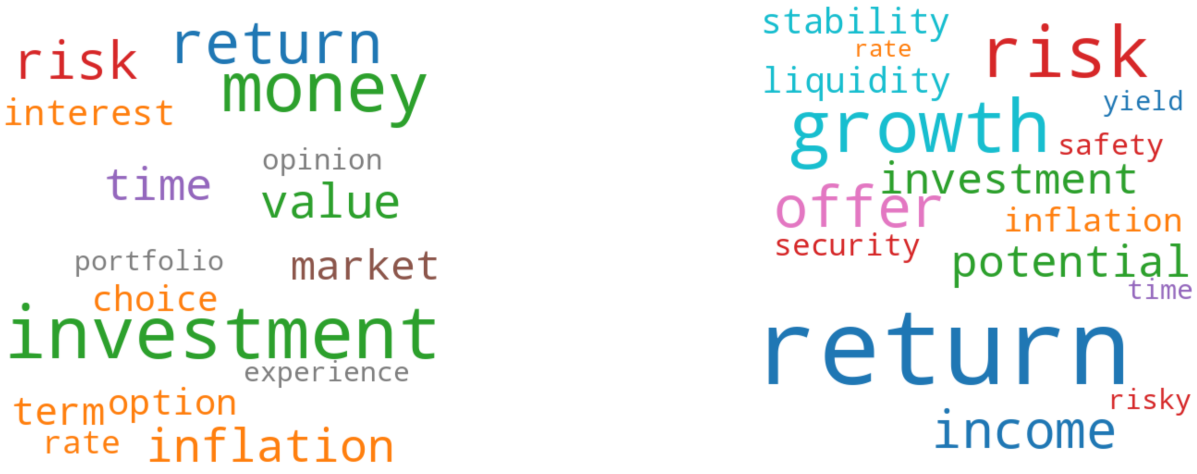
We now examine whether the capabilities of generative AI extend beyond matching investment ratings to the reasoning behind them. Specifically, we analyze the free-form justifications accompanying each rating to assess the extent to which large language models replicate the underlying structure of human explanations.

4.1 Unigram analysis: Most common themes

We begin by identifying the most common themes discussed in human and AI-generated responses. Figure 2 offers a graphical illustration of a simple word count of all nouns in the human survey responses (on the left) and AI-generated responses (on the right).[12] Human responses use a more varied language than the AI model, but the most common

[12] Table A2 in the Online Appendix shows the 15 most common nouns in each dataset with corresponding frequencies.

themes in both sets of responses are similar, focusing on investments, risk, and return. There are differences in the precise terms used—for example, "investment" and "money" in human responses versus "return" and "growth" in AI-generated responses—but the general patterns are similar: both sets of explanations concentrate on financial tradeoffs.



(a) Human word cloud

(b) AI word cloud

Figure 2: Word clouds of nouns in human (left-hand panel) and AI-generated (right-hand panel) free-form explanations. The relative size of each word in the word cloud reflects that word's frequency.

4.2 Semantic embedding: Risk and return

Motivated by the themes emerging from the most frequent terms, we conduct a more rigorous examination of the two main themes in the responses, risk and return. We evaluate how explanations accompanying higher versus lower ratings of different investment options differentially refer to risk and return as rationales, and whether these relationships differ between human and AI-generated responses.

To conduct this analysis, we construct numerical representations of the relevant text in the free-form responses, so-called embeddings. Specifically, our approach builds on the idea of Semantic Axis (SemAxis) in the natural language processing literature, which uses differences in embeddings of words in opposite semantic classes (e.g., happy vs. sad) to build a numerical scale of a meaning (An, Kwak and Ahn, 2018). Applying this approach to our setting, we first construct two semantic dimensions, a risk dimension and a return dimension, based on the embeddings of the following four sentences:

Return: "This asset has very high return." *versus* "This asset has very low return.";
Risk: "This asset has very high risk." *versus* "This asset has very low risk."

We denote the embeddings of these four sentences as vectors $\mathbf{V}^h_{\text{ret}}$, $\mathbf{V}^l_{\text{ret}}$, $\mathbf{V}^h_{\text{risk}}$, and $\mathbf{V}^l_{\text{risk}}$, respectively (normalized to have unit length). By design, the only conceptual difference between $\mathbf{V}^h_{\text{ret}}$ and $\mathbf{V}^l_{\text{ret}}$ is in the return dimension: high versus low. Therefore, by taking the difference between the vectors $\mathbf{V}^h_{\text{ret}}$ and $\mathbf{V}^l_{\text{ret}}$, we obtain a vector that defines the axis pointing from low to high returns:

$$\mathbf{V}_{\text{ret}} = \frac{\mathbf{V}^h_{\text{ret}} - \mathbf{V}^l_{\text{ret}}}{\text{norm}\left(\mathbf{V}^h_{\text{ret}} - \mathbf{V}^l_{\text{ret}}\right)}. \tag{1}$$

Similarly, we obtain a vector axis pointing from low to high risk by differencing $\mathbf{V}^h_{\text{risk}}$ and $\mathbf{V}^l_{\text{risk}}$:

$$\mathbf{V}_{\text{risk}} = \frac{\mathbf{V}^h_{\text{risk}} - \mathbf{V}^l_{\text{risk}}}{\text{norm}\left(\mathbf{V}^h_{\text{risk}} - \mathbf{V}^l_{\text{risk}}\right)}. \tag{2}$$

We note that the two axes $\mathbf{V}_{\text{ret}}$ and $\mathbf{V}_{\text{risk}}$ are nearly orthogonal, with an angle of 70 degrees.

Next, to measure the risk and return component of each individual free-form explanation i, we extract the embedding vector $\mathbf{emb}_i$ of i and then identify its return-related component $c_{i,r}$ and risk-related component $c_{i,v}$ as:

$$\begin{aligned} \mathrm{c}_{i,r} &= \cos(\mathbf{emb}_i, \mathbf{V}_{\text{ret}}), \\ \mathrm{c}_{i,v} &= \cos(\mathbf{emb}_i, \mathbf{V}_{\text{risk}}), \end{aligned} \tag{3}$$

where cos denotes cosine similarity between two vectors. Intuitively, $c_{i,r}$ is explanation i's association with high return, and $c_{i,v}$ is i's association with high risk. We do this for each human and AI-generated response.[13]

Using these components, we compare the relative ordering of the asset classes (stocks, bonds, and cash) along the return and risk dimensions between human and AI-generated responses. As shown in Table 5, both human participants' and AI-generated explanations have, on average, the highest return components when explaining ratings of stocks, followed by bonds, then cash. Similarly, both human and AI-generated explanations project

[13] Some examples of explanations from our data with high and low risk and return components are shown in Table A3 in the Online Appendix.

	Risk and Return Ranking:			
	Risk		Return	
	Human	AI	Human	AI
highest	stock	stock	stock	stock
middle	cash	cash	bond	bond
lowest	bond	bond	cash	cash

Table 5: This table shows the implicit (based on the embeddings of the free-form explanations) rankings of the three asset classes—stocks, bonds, and cash—along the risk and return dimensions, sepatately for human and AI-generated responses.

the highest risk when discussing stocks, followed by cash, and then bonds. Thus, AI's relative discussion of stocks, bonds, and cash along both risk and return axes is consistent with the discussion provided by human survey participants.

Next, we examine how risk and return in the free-form explanations relate to the numerical ratings of the corresponding investment options, and whether this relationship differs between human and AI-generated responses. This analysis accomplishes two goals: (i) to observe the extent to which the categorical ratings are explained by risk and return considerations, and (ii) to test whether the relationship between categorical ratings and free-form discussions of risk and return in actual human survey responses is correctly reflected in AI-generated data. We estimate the following specification:

$$\begin{aligned}\text{Rating}_{i,k} =& \beta_1 c_{i,k,r} + \beta_2 c_{i,k,v} + \beta_3 \mathbb{1}(\text{i is an AI agent}) * c_{i,k,r} \\ &+ \beta_4 \mathbb{1}(\text{i is an AI agent}) * c_{i,k,v} + \delta_{k,AI} + \epsilon_{i,k}, \end{aligned} \quad (4)$$

where $\text{Rating}_{i,k}$ is participant i's rating of investment option $k \in \{\text{stocks}, \text{bonds}, \text{cash}\}$, standardized to have a mean of zero and a standard deviation of one. The independent variables $c_{i,k,r}$ and $c_{i,k,v}$ are the projections of participant i's free-form responses regarding investment option k onto the return (r) and risk (v) dimensions, likewise standardized to have a mean of zero and a standard deviation of one. The coefficient β_1 (β_2) captures the importance of return (risk) for human investment ratings. Correspondingly, the interaction-effect coefficent β_3 (β_4) reflects the under- or overstatement of the importance of return (risk) by generative AI, relative to human rationales. Finally, $\delta_{i,k,AI}$ are fixed effects of investment options k interacted with an indicator for AI-generated data.

The results are presented in column (1) of Table 6. The *return* coefficient indicates that

	Dependent variable:	
	Rating	
return	0.788^{***}	0.789^{***}
	(0.015)	(0.015)
risk	-0.564^{***}	-0.549^{***}
	(0.017)	(0.017)
AI × return	-0.010	-0.029
	(0.021)	(0.021)
AI × risk	0.003	0.011
	(0.028)	(0.028)
Asset class × AI	✓	✓
Demographics × AI		✓
Observations	6,348	6,348
R^2	0.556	0.560

Table 6: The associations between risk and return projections of free-form explanations and the corresponding investment ratings in the pooled human and AI-generated data. The left column includes fixed effects for the three asset classes (stocks, bonds, cash) interacted with an AI indicator. The right column also controls for demographics (age, gender, income).

higher perceived return-generating potential of an asset in free-form responses is significantly positively associated with higher categorical ratings for that asset. Importantly, this relationship is statistically indistinguishable between human and AI-generated responses, as indicated by the small and insignificant coefficient on the interaction term *AI x return*. The estimates imply that a one-standard-deviation increase in the perceived return is associated with a 0.79-standard-deviation increase in the rating in the human survey and a 0.78-standard-deviation increase in the rating in AI-generated responses, with the two estimates being statistically indistinguishable.

A similarly aligned pattern is observed on the risk side: a one-standard-deviation increase in the risk projection is associated with a 0.56-standard-deviation decrease in the rating in both human and AI-generated data.

Column (2) of Table 6 shows that these results are robust to including demographic characteristics interacted with the data source when estimating equation (4). Overall, we observe that AI-generated responses closely match human responses in terms of the "reasoning" about risk-return considerations that drive investment ratings.

4.3 The roles of knowledge and past experiences

We now investigate the other major rationales, besides risk and return, that play a role in the data, and the extent to which the references to these themes align between human and AI-generated free-form responses. We focus on explanations for stock ratings, to speak to the drivers of low stock market participation, which is a long-standing and significant issue documented in the prior literature (Guiso, Sapienza and Zingales, 2008; Van Rooij, Lusardi and Alessie, 2011).[14]

For this analysis, we combine *all* of the explanations about investing in stocks from human survey participants and divide these explanations into two subsets: explanations associated with negative ratings (≤ 2) and explanations associated with positive ratings (≥ 4). We then use generative AI's summarization capabilities to answer the following prompt based on the two subsets of positive/negative explanations:

> *Read the following two sets of opinions about investing in stocks and describe 5 themes other than risk and return that are different between the two sets using 5 short phrases:*
>
> *set 1:... (explanations of positive ratings)*
>
> *set 2:... (explanations of negative ratings)*

These queries produce the five main themes (other than risk and return) in the human response data. We use the same prompt to identify the main themes in the responses generated by AI. We repeat the procedure three times to ensure that we obtain consistent, replicable summaries and do not capture statistical noise. The full list of all themes from each run is listed in Table A4 in the Online Appendix.

The results reveal that the most prominent topics in both human and AI-generated data are "knowledge" (ease of understanding the stock market) and "experiences" (positive or negative emotional reactions to past experiences with the stock market). As visualized in Figure 3, the themes related to knowledge and understanding in the human data are summarized by "Perception of Complexity and Accessibility" (first run), "Understanding and Accessibility" (second run), and "Perception of Complexity" (third run). Similarly, AI-generated responses reflect the theme "Knowledge and Complexity" (second and third run). And the themes related to "past experiences" in the human

[14] In untabulated analyses, we examine the auxiliary themes for bonds, finding that knowledge/familiarity emerges as the main auxiliary theme, and cash, finding that the most consistent theme is the convenience/accessibility of cash.

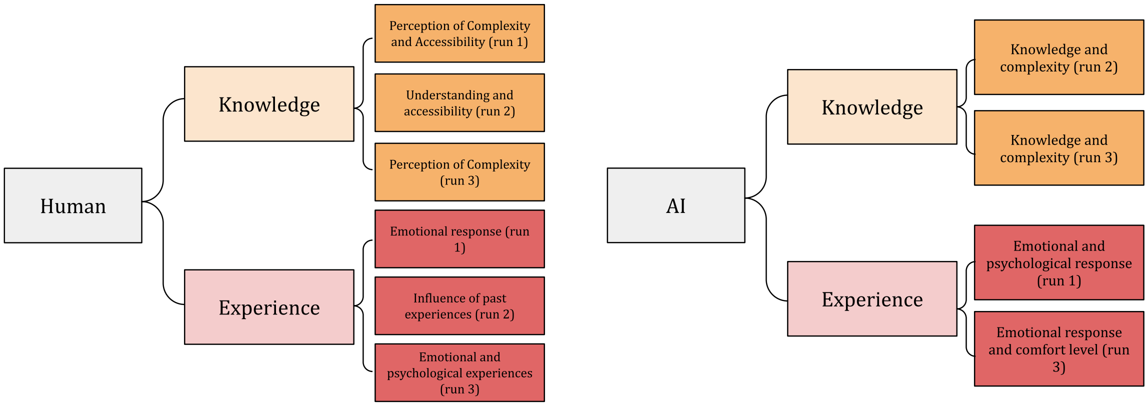

Human
Knowledge
Perception of Complexity and Accessibility (run 1)
Understanding and accessibility (run 2)
Perception of Complexity (run 3)
Experience
Emotional response (run 1)
Influence of past experiences (run 2)
Emotional and psychological experiences (run 3)
AI
Knowledge
Knowledge and complexity (run 2)
Knowledge and complexity (run 3)
Experience
Emotional and psychological response (run 1)
Emotional response and comfort level (run 3)

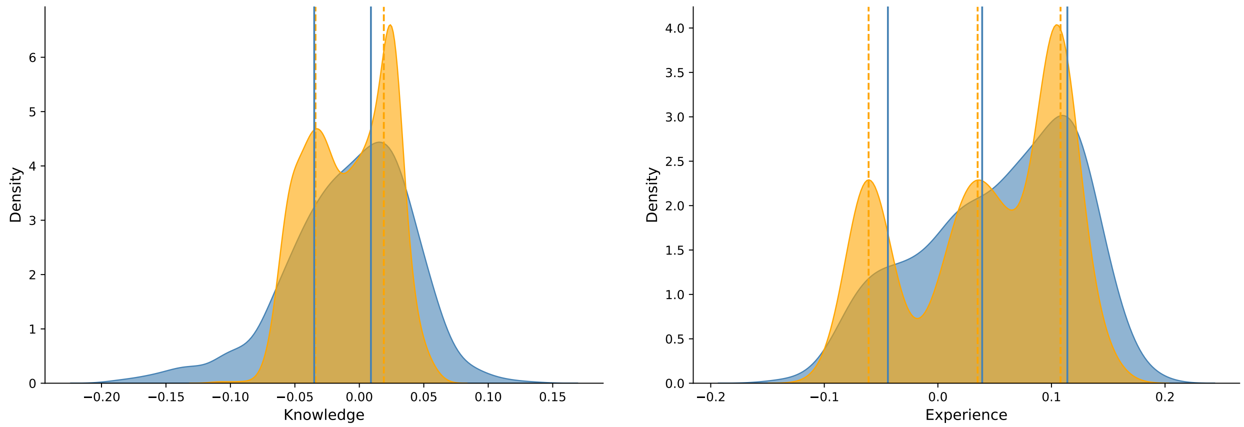
Figure 3: This figure shows the auxiliary themes in free-form explanations of stock ratings, extracted using AI. The themes summarize the main differences (other than risk and return) between explanations accompanying high (4–5) versus low (1–2) stock ratings. The themes extracted from human survey responses are shown on the left, and those extracted from AI-generated responses are on the right.

survey responses include “Emotional Response” (first run), “Influence of Past Experiences” (second run), and “Emotional and Psychological Experiences” (third run), while AI-generated responses include “Emotional and Psychological Response” (first run) and “Emotional Response and Comfort Level” (second run).

Next, we examine whether AI-generated responses capture the heterogeneous prevalence of the knowledge and experience rationales among different demographic groups. We use an approach similar to the analysis of the risk and return dimensions to quantify the reliance on “knowledge” and “past experiences” in the free-form explanations. We relate these measures to age, gender, and income of the human or AI-generated respondent and test whether the heterogeneity among human respondents is mirrored in the AI-generated data.

We first consider the knowledge dimension. Using the Semantic Axis approach discussed in Section 4.1, we define the knowledge dimension as the difference between the embeddings of the following two sentences: “I am very knowledgeable about the stock market” (high knowledge), and “I do not know anything about the stock market” (low knowledge). Then, we project the embedding of each explanation accompanying a given (human or AI-generated) rating of stocks onto this knowledge dimension.

Figure 4(a) shows a graphical representation of this projection, in blue for the explanations provided by human participants and in yellow for the explanations generated



(a) Distribution of knowledge projections.

(b) Distribution of past experiences.

Figure 4: The density of human and AI-generated responses' loadings on having a high level of knowledge about the stock market (subfigure a) and positive versus negative experiences with the stock market (subfigure b). The blue density plot represents the human data, and the yellow plot represents the AI-generated data. Continuous vertical lines mark the centers of the clusters of human data, and dotted vertical lines mark the centers of the clusters of AI-generated data.

by AI. Both projections center just below zero, and the tails of both density distributions stretch from the negative to the positive realm. AI-generated responses show a clear bimodal distribution and smaller tails, while human responses are less bimodal and have longer tails.

In order to compare the distributions, we group data points into a positive cluster, which reflects the presence of knowledge regarding stocks, and a negative cluster, which reflects the absence of knowledge or understanding. We identify the clusters using a Gaussian Mixture Model, modeling each set of knowledge projections (from human and AI-generated explanations) using a data-generating process that randomly picks individuals from two Gaussian distributions (one for each cluster) with some fixed probabilities as follows:

1. Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_m)$ be the initial means of the m clusters, $\boldsymbol{\Sigma} = (\Sigma_1, \Sigma_2, \cdots, \Sigma_m)$ be the initial covariance matrices, and $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_m)$ be the initial mixing coefficients. In the case of the knowledge dimension, we set the number of clusters $m = 2$.

2. Calculate the responsibility r_{ik} for each data point i and each cluster k using the

current parameter estimates:

$$r_{ik} = \frac{\pi_k \phi_k(\mathbf{x_i})}{\sum_{j=1}^{m} \pi_j \phi_j(\mathbf{x_i})},$$

where ϕs are the PDFs of the estimated multivariate Gaussian distributions.

3. Update the parameters:

$$\begin{aligned}
\boldsymbol{\mu}_k^{new} &= \frac{\sum_{i=1}^{N} r_{ik} \mathbf{x}_i}{\sum_{i=1}^{N} r_{ik}}, \\
\boldsymbol{\Sigma}_k^{new} &= \frac{\sum_{i=1}^{N} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k^{new})(\mathbf{x}_i - \boldsymbol{\mu}_k^{new})^T}{\sum_{i=1}^{N} r_{ik}}, \\
\pi_k^{new} &= \frac{1}{N} \sum_{i=1}^{N} r_{ik}.
\end{aligned}$$

4. Repeat the above two steps until convergence.[15]

The optimizations converge with fewer than 100 iterations. The resulting clusters are displayed in Figure 4(a), with solid blue vertical lines for human survey explanations and dashed yellow vertical lines for AI-generated explanations. Although AI-generated explanations tend to be slightly more positive, the overall clustering pattern remains similar: Cluster 1, which includes explanations reflecting limited knowledge of the stock market, is centered around -0.035 (-0.034) for human (AI-generated) responses; Cluster 2, containing explanations that demonstrate a high level of market knowledge, is centered around 0.009 (0.019) for human (AI-generated) responses.

Next, we examine how knowledge levels vary across demographic characteristics, and whether these patterns are correctly reflected by AI. For this step of the analysis, we classify each individual embedding into either Cluster 1 (respondents who do not feel confident about the stock market or think it is too complex) or Cluster 2 (respondents who feel knowledgeable about the stock market). For each data set (embeddings of human responses and embeddings of AI-generated responses), we regress the cluster label on the following three demographic characteristics: age, gender, and income.

[15] We use a convergence tolerance of 0.0001. For more details about Gaussian Mixture Models and other finite mixture models, refer to McLachlan, Lee and Rathnayake (2019).

	Dependent variable:			
	Knowledge		Experience	
	Human	AI	Human	AI
age	−0.001	−0.015***	−0.004**	−0.027***
	(0.001)	(0.002)	(0.002)	(0.003)
gender	−0.092***	−0.079***	−0.201***	−0.095**
	(0.025)	(0.027)	(0.047)	(0.041)
income	0.002***	0.057***	0.003***	0.110***
	(0.0004)	(0.003)	(0.001)	(0.005)
Observations	1,074	1,042	1,074	1,042
R^2	0.048	0.239	0.039	0.334

Table 7: This table shows the associations between stock market knowledge (columns 1 and 2) and past experiences (columns 3 and 4) and demographic characteristics (age, gender, and income). Columns 1 and 3 present the results for human responses, while columns 2 and 4 present the results for AI-generated responses. Income is in thousands of dollars. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

The first two columns of Table 7 show the results. In both human and AI-generated data, free-form responses of men reflect a higher level of knowledge about the stock market than the free-form responses of women (with statistically indistinguishable coefficient estimates between human and AI-generated data). The same holds for human and AI-generated responses of high- versus low-income individuals (albeit with a more substantial difference in coefficient estimates). In AI-generated data, there is also a negative association between age and expressed knowledge of the stock market, which is not significant in human data.

We next conduct a similar analysis of the "experience" and "emotional response" dimension, using the difference between the embeddings of the following two sentences: "I have had very good experiences investing in the stock market." (positive experiences) and "I have had terrible experiences investing in the stock market." (negative experiences).

Figure 4(b) shows the resulting projections, and we apply the same approach as we did for the knowledge theme (i.e., a Gaussian Mixture Model) to cluster these distributions, defining three clusters corresponding to more positive experiences, more negative experiences, and relatively neutral experiences. Solid blue vertical lines in Figure 4(b) mark the centers of the three clusters for the experience projections of human responses, and dotted yellow vertical lines represent the centers of the three clusters for the projections of AI-generated responses. The cluster of negative past experiences is centered around -0.044 (-0.061) in the human (AI-generated) data, the cluster of neutral to mildly positive experiences is centered around 0.039 (0.035), and the cluster of strongly positive experiences is centered around 0.114 (0.108). The distribution of responses across these clusters is similar between human and AI-generated data: 23% (24%) of human (AI-generated) responses fall into the negative cluster, 33% (35%) fall into the neutral to mildly positive cluster, and 44% (41%) fall into the strongly positive cluster.

The last two columns of Table 7 examine how reported past experiences with the stock market vary across demographics, and whether AI captures those differences. We regress the cluster labels on age, gender, and income of the corresponding human participant or simulated AI agent. In human data, younger individuals tend to have more positive experiences with the stock market than older individuals, men tend to have more positive experiences than women, and high-income individuals tend to have more positive experiences than low-income individuals. All coefficients are significant at the 5% level or more. These directional patterns are the same in AI-generated data: young, male, and high-income AI agents generate explanations reflecting more positive experiences with the stock market, statistically significant at the 5% level or more.

Altogether, Table 7 shows that there are strong demographic patterns in the expressions of both knowledge and past experiences of the stock market, and that generative AI is effective in capturing these directional patterns across demographic groups.

4.4 Text-revealed factor structure

Our analysis shows that risk, return, knowledge, and past experiences are the key themes shaping attitudes towards the stock market. Generative AI directionally replicates human patterns in each of these dimensions. Next, we examine how these themes relate to each other and whether generative AI accurately captures their interconnections.

Table 8 presents the associations between the four themes in the *human* investment survey, controlling for demographic characteristics (age, gender, and income). Both knowledge and positive past experiences are associated with higher return expectations and lower risk perceptions. Moreover, knowledge of the stock market is strongly associated with positive past experiences (coefficient of 0.500).

Table 9 replicates this analysis for AI-generated responses, revealing similarly strong and statistically significant associations. For example, knowledge about the stock market exhibits an association of 0.398 with positive past experiences. These results suggest that generative AI captures not only individual themes but also the relationships between them, closely mirroring the associations observed in human explanations.

Our results indicate that generative AI can replicate not only investment ratings and their underlying justifications but also the structural relationships between these justifications, aligning with the patterns found in human survey data. This was ex-ante an open question given the functioning of generative AI models. These models are trained on vast amounts of textual data to generate contextually appropriate responses. While this allows them to produce accurate average ratings and capture isolated themes, it does not necessarily guarantee the correct co-occurrence of themes within responses.

	Dependent variable:				
	return		risk		knowledge
experience	0.636^{***}		-0.312^{***}		0.500^{***}
	(0.016)		(0.016)		(0.014)
knowledge		0.919^{***}		-0.184^{***}	
		(0.023)		(0.027)	
Observations	1,074	1,074	1,074	1,074	1,074
R^2	0.627	0.619	0.260	0.052	0.556

Table 8: This table shows the associations between the four main themes in free-form explanations of human survey participants' stock ratings: risk, return, experience, and knowledge. The associations are computed controlling for participants' gender, age, and income. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

5 Transitivity violations and lack of financial knowledge

The ability of generative AI models to capture heterogeneous investment preferences *and* the rationales underlying these heterogeneous choices suggests that AI has the potential to represent a broad investor base. Furthermore, the ability of generative AI to capture the *rationales* behind the expressed preferences indicates promise of AI models to eventually help identify and screen out biases in investing.

In the final part of the paper, we focus on one such bias, violations of transitivity. We provide evidence of non-transitive preference orderings among human survey respon-

	Dependent variable:				
	return		risk		knowledge
experience	0.815^{***}		-0.202^{***}		0.398^{***}
	(0.014)		(0.016)		(0.008)
knowledge		1.584^{***}		-0.328^{***}	
		(0.037)		(0.036)	
Observations	1,042	1,042	1,042	1,042	1,042
R^2	0.848	0.762	0.153	0.110	0.100

Table 9: This table shows the associations between the four main themes in free-form explanations of AI-generated stock ratings: risk, return, experience, and knowledge. The associations are computed controlling for gender, age, and income. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

dents and utilize textual analysis to illustrate how AI can help identify where investment choices are hampered by the lack of financial knowledge.

5.1 Violations of transitivity

In economic theory, transitivity is a necessary condition for preferences to be deemed rational. We first investigate whether this condition holds for human and AI-generated responses. For this analysis, we leverage the third part of our survey experiment, where we elicit direct comparisons between stocks and bonds, bonds and cash, and cash and stocks.

To test whether respondents' preference orderings satisfy transitivity, we use the following proposition:

Proposition 1. *Let $a \in \{stocks, bonds, indifferent\}$, $b \in \{stocks, cash, indifferent\}$, and $c \in \{cash, bonds, indifferent\}$ be the three responses provided by a survey participant to the three relative comparison questions. The participant's preference ordering satisfies transitivity if and only if the following conditions are met:*

i. If $a, b, c \neq indifferent$, exactly one of the following must be true: $a = b$, $b = c$, or $a = c$.

ii. If $a = indifferent$, then either $b = c$ or $b, c \in \{stocks, bonds\}$.

iii. If $b = indifferent$, then either $a = c$ or $a, c \in \{cash, stocks\}$.

iv. If $c = indifferent$, then either $a = b$ or $a, b \in \{cash, bonds\}$.

Using this proposition, we calculate the frequencies of transitive versus non-transitive preference orderings. As shown in Table 10, we find that 84.4% of human survey participants and 98.7% of simulated AI agents follow transitivity. Thus, AI agents' preferences are almost always transitive, while human responses violate transitivity over 15% of the time.

In the next step, we aim to identify the source(s) of transitivity violations in human survey data, in contrast to the negligible non-transitivity rate in AI-generated data. We start by dividing the human sample into responses from men and women. As shown in Table 10, 89.5% of male survey participants have transitive preferences, compared to only 79.1% of female survey participants. We then split the sample of human participants into those who give strict orderings (no ”indifferent” responses) and those who have at least one ”indifferent” response. 95.7% of participants *without* ”indifferent” responses have transitive preferences, compared to only 72.0% of participants with at least one ”indifferent” response.[16]

Combining these two observations, we investigate whether the lower proportion of female participants with transitive preference orderings is due to the higher incidence of indifference among women. Indeed, we find that 40.3% of male survey participants have at least one ”indifferent” response, compared to 55.7% of female survey participants. In addition, we observe that conditioning on the human participants who responded with at least one indifference, men are still more likely to have transitive preferences than women: 79.0% of men compared to 66.8% of women. Among participants with no ”indifferent” responses, the share of transitive preference orderings is not statistically different between men (96.6%) and women (94.5%).

This analysis reveals three novel insights related to the rationality of expressed preference orderings. First, generative AI agents almost always have transitive preference orderings, suggesting that transitivity is one area in which AI can help debias human responses. Second, among human survey participants, women are more likely to violate transitivity than men. Third, the source of this gender discrepancy is expressed indifference between investment options: women are more likely to express indifference in the survey; and conditional on at least one ”indifferent” response, women's weak preference

[16] Note that no simulated AI agent responded with “indifferent” despite being provided this option. This is potentially because AI is trained to produce answers that are preferred by human requesters, and when a human asks AI to make a selection, “indifferent” is usually not a desirable answer.

	AI		Human			
Observations	1,042		1,074			
Transitive prob	98.7%		84.4%			
Difference	14.3%***					
	Male	Female	Male		Female	
Observations	500	542	544		530	
Transitive prob	99.8%	97.6%	89.5%		79.1%	
Difference	2.2%***		10.4%***			
			diff	indiff	diff	indiff
Observations			325	219	235	295
Transitive prob			96.6%	79.0%	94.5%	66.8%
Difference			17.6%***		27.7%***	

Table 10: Preference transitivity of human participants and simulated AI agents. In this table, the *diff* columns report the results from the subset of human participants who did not respond with "indifferent" to any of the three preference questions. The *indiff* columns report the results from the subset of human participants who responded with "indifferent" to at least one of the three preference questions. The rows denoted *Transitive prob* report the frequencies of agents' answers satisfying transitivity. *Difference* indicates the difference in frequencies between either male and female respondents or between those who never respond with indifference and those who indicated at least one indifference. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

orderings are more likely to violate transitivity than men's.

5.2 Financial knowledge and indifference towards investment options

We now turn to the free-form explanations from the human survey data for each set of pairwise comparisons—stocks versus bonds, bonds versus cash, and cash versus stocks—to uncover the reasons underlying expressed indifference and non-transitivity.

Similar to Section 4, we use AI to summarize the themes that differentiate two subsets of the free-form data: human participants who responded with at least one indifference and those who provided only strict preferences. We use the following prompt:

> *Read the following two sets of opinions about investing in stocks, bonds, and cash and describe 5 themes other than risk and return that are different between the two sets using 5 short phrases:*
>
> *set 1:... (explanations without indifference.)*

> *set 2:... (explanations with indifference.)*
>
> *Phrase them in pairs of prompts such as: "investments have high risk" versus "investments have low risk."*

We perform the summarization exercise three times using a temperature of 0.8.

As shown in Online Appendix Table A5, the main topic that consistently appears among the top five themes is knowledge about bonds and stocks. The first run of the summarization exercise features "asset familiarity" and "investment understanding" among the top five topics, the second run features "familiarity and understanding," and the third run features "familiarity and knowledge."

With this finding in hand, we turn to the individual free-form responses and measure the level of knowledge reflected in each participant's response. We use the following projection approach, similar to the one discussed in Section 4: We first construct the semantic axis using a pair of prompts: "I am knowledgeable about investing in stocks and bonds" (high knowledge) and "I know little about investing in stocks and bonds" (low knowledge). We denote the embeddings of these two sentences as $\mathbf{V}^h_{\text{knowledge}}$ and $\mathbf{V}^l_{\text{knowledge}}$, respectively. Then we combine the text explanations of each participant i into one string, take its embedding $\mathbf{emb}_i$, and estimate participant i's knowledge level as

$$c_{i,\text{knowledge}} = \cos\left(\mathbf{emb}_i, \mathbf{V}_{\text{knowledge}}\right),$$

where $\mathbf{V}_{\text{knowledge}} = \frac{\mathbf{V}^h_{\text{knowledge}} - \mathbf{V}^l_{\text{knowledge}}}{\text{norm}\left(\mathbf{V}^h_{\text{knowledge}} - \mathbf{V}^l_{\text{knowledge}}\right)}$. We then standardize $c_{i,\text{knowledge}}$ to a mean of zero and a standard deviation of one.

We estimate the association between knowledge and indifference using a logistic regression:

$$P(\text{indiff}_i = 1) = \frac{1}{1 + e^{-(\alpha_{\text{knowledge}} + \beta_{\text{knowledge}} c_{i,\text{knowledge}} + \text{controls})}}, \tag{5}$$

where $\textit{indiff}_i$ is a binary indicator for whether participant i has expressed at least one indifference between two investment options. We estimate this regression with three types of controls: no controls, demographic controls, and demographic controls interacted with the knowledge projection. The results are presented in Table 11.

We find that, across all specifications, greater knowledge about stocks and bonds is significantly negatively associated with a participant being indifferent about at least two investment options. A one-standard-deviation increase in the knowledge dimension decreases the log-odds of being indifferent by approximately 0.72 to 0.79. Thus, participants who are more knowledgeable about stocks and bonds are much less likely to express indifference between two or more options. Moreover, comparing the t-statistics and McFadden (1974) pseudo-R^2, we see that self-reported knowledge about stocks and bonds in the free-form explanations has a stronger association with indifference than demographic variables, and the interaction of knowledge with demographics adds little incremental explanatory power.

	Dependent variable:		
	indiff		
Knowledge	-0.773^{***} (0.075)	-0.720^{***} (0.076)	-0.790^{***} (0.277)
Demographics		✓	✓
Demographics × knowledge			✓
Observations	1,074	1,074	1,074
McFadden R^2	0.0871	0.1139	0.1158

Table 11: This table shows the association between the level of knowledge reflected in each participant's free-form explanations accompanying comparison questions and whether the participant is indifferent between at least one pair of investment options. *Knowledge* is standardized to have a mean of zero and a standard deviation of one. The left column has knowledge as the only independent variable. The center column adds demographic variables (age, gender, and income) as additional controls. The right column adds demographic variables interacted with knowledge as additional controls. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Next, we examine the association between expressed knowledge and transitive preference orderings, with a particular focus on whether this association is mediated by indifference towards at least two investment options. We adopt the mediation analysis approach detailed in Tingley et al. (2014). To ensure proper numerical representation, the knowledge variable is transformed into a binary indicator, where participants with knowledge above the median are coded as one and those below the median are coded as zero.

The mediation analysis proceeds in four steps: fitting a mediator model, fitting a dependent variable model, estimating direct and mediated associations, and bootstrapping the confidence intervals.

For the first step, we estimate the following model for the mediator (indifference):

$$M(t_i, X_i) = P(I_i = 1) = \frac{1}{1 + e^{-\alpha_0^I + \beta_1^I t_i + \beta_2^I X_i}},$$

where t_i captures the knowledge level (0 or 1) and X_i denotes demographic controls (gender, age, and income). $M(t_i, X_i)$ is the estimated probability of participant i reporting at least one indifference between two investment options ($I_i = 1$).

The second step involves fitting the model for the dependent variable of interest ("having a transitive preference ordering"):

$$Y(t_i, I_i, X_i) = P(Tr_i = 1) = \frac{1}{1 + e^{-\alpha_0^{Tr} + \beta_1^{Tr} t_i + \beta_2^{Tr} I_i + \beta_3^{Tr} X_i}},$$

where Tr_i is an indicator for participant i having a transitive preference ordering.

Third, we compute the mediated and direct associations using the models estimated above. The mediated association captures the expected difference in the probability that participant i exhibits a transitive preference ordering when the mediator (i.e., reporting at least one indifference towards investment options) changes in response to the independent variable of interest (i.e., the participant's knowledge), while the independent variable itself remains fixed. This association is expressed as

$$\delta_i(t) = Y(t, M(t_1, X_i), X_i) - Y(t, M(t_0, X_i), X_i).$$

The direct association represents the expected difference in the probability that participant i has transitive preference orderings when his or her financial knowledge level varies, while holding constant the mediator (probability of reporting at least one indifference between investment options) at the value it would take under a specific knowledge level t. This is expressed as

$$\zeta_i(t) = Y(t_1, M(t, X_i), X_i) - Y(t_0, M(t, X_i), X_i).$$

We use a bootstrap approach with 1.000 simulations to calculate confidence intervals for the mediated and direct associations. For each bootstrap sample, we refit the models M and Y and compute the average direct and mediated association between participants' knowledge level and preference transitivity over the entire sample of participants. The estimates from these 1,000 runs allow us to construct 95% confidence intervals.

The mediated association between participants' knowledge level and preference transitivity is estimated at 0.059 and statistically significant at the 1% level, indicating that, on average, having above-median knowledge is associated with a 0.059 log-odds increase in transitive preferences. In contrast, the direct association between knowledge and preference transitivity is not significant (with a coefficient of -0.016 and a p-value of 0.496).

In summary, our analysis demonstrates that participants' knowledge about investing in stocks and bonds is significantly positively associated with participants expressing preference orderings that are strict (without indifference) and transitive. Importantly, the relationship between knowledge and transitive preferences is primarily mediated by indifference—thus, human survey participants appear to use indifference as a way to express uncertainty or a lack of a clear coherent preference ordering.

6 Conclusion

With the rise of machine learning and artificial intelligence, algorithmic bias is becoming an increasingly salient issue, including in finance. For example, banks have faced lawsuits for using AI models that lead to discriminatory lending, and Bartlett et al. (2022) showcases multiple concerns in the credit space.

We examine how well generative AI can replicate human investment preferences, not only on average but also across demographics. We show that the default AI model overrepresents young high-income individuals. However, when prompted to assume specific demographic identities, generative AI correctly captures preference heterogeneity across gender, age, and income—in both investment preferences and actual asset allocations. Furthermore, generative AI replicates the rationales behind those heterogeneous preferences, as reflected in free-form explanations. Finally, we leverage AI to study the cause of transitivity violations in human survey participants, which are not inherited by AI. We show that violations of transitivity arise from a lack of financial knowledge, which human respondents express as indifference between at least two investment options.

These results lay the foundation for two potential directions for future research. First, the rising prominence of robo-advising in the investment management space (D'Acunto, Prabhala and Rossi, 2019; Rossi and Utkus, 2021) raises the question of how the advent of generative AI may affect different investor groups. Good financial advice needs to accomplish two goals: (i) understand the clients (existing preferences, constraints, limitations, etc.), and (ii) guide the clients (towards better allocations, more diversified portfolios, etc.). Our paper shows that generative AI does well on the first goal, ”understanding” human investment preferences, including the rationales behind them. Future research can build on this finding by examining the ability of generative AI to guide heterogeneous human investors towards better allocations. As the investment landscape becomes not only more automated but also more diverse, with growing shares of small and retail investors, the ability to accurately reflect cross-demographic heterogeneity in investment preferences, attitudes, and experiences will be crucial for successful applications of AI in finance.

Second, the ability of generative AI to capture the rationales in free-form explanations suggests the promise of uncovering mechanisms behind survey responses and elicited preferences. Stantcheva (2020), Ferrario and Stantcheva (2022), and Haaland et al. (2024) highlight the importance of including open-ended questions in surveys to better understand people’s expectations and reactions to economic policies. Our results offer a methodological innovation: generative AI can be used to replicate the full survey design, from categorical (closed-ended) questions to open-ended explanations, preserving the factor structure of the underlying themes. This is especially important in a domain such as investment preferences, with substantial heterogeneity in preferences and reasoning. When differences in preferences are driven by different risk aversion parameters, investment advisers should take this into account. When, instead, differences in preferences are due to incorrect information or lack of financial literacy, the role of investment advice is to inform or debias. Our paper makes the first step in showcasing that generative AI can be a powerful tool for uncovering underlying mechanisms, potentially reducing the need to run costly and time-consuming human surveys.

References

Agnew, Julie, Pierluigi Balduzzi, and Annika Sunden. 2003. “Portfolio choice and trading in a large 401 (k) plan.” *American Economic Review*, 93(1): 193–215.

An, Jisun, Haewoon Kwak, and Yong-Yeol Ahn. 2018. “SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment.” 2450–2461. Melbourne, Australia:Association for Computational Linguistics.

Araci, D. 2019. “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.” *arXiv preprint arXiv:1908.10063.*

Barber, Brad M, and Terrance Odean. 2001. "Boys will be boys: Gender, overconfidence, and common stock investment." *The Quarterly Journal of Economics*, 116(1): 261–292.

Barber, Brad M, Xing Huang, Terrance Odean, and Christopher Schwarz. 2022. "Attention-induced trading and returns: Evidence from Robinhood users." *The Journal of Finance*, 77(6): 3141–3190.

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace. 2022. "Consumer-lending discrimination in the FinTech era." *Journal of Financial Economics*, 143(1): 30–56.

Bertomeu, Jeremy, Yupeng Lin, Yibin Liu, and Zhenghui Ni. 2023. "Capital Market Consequences of Generative AI: Early Evidence from the Ban of ChatGPT in Italy." *Available at SSRN 4452670*.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. "Language models are few-shot learners." *Advances in neural information processing systems*, 33: 1877–1901.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond. 2023. "Generative AI at Work." National Bureau of Economic Research.

Bybee, J Leland. 2023. "The Ghost in the Machine: Generating Beliefs with Large Language Models." Yale University Working Paper.

Cao, Sean, Wei Jiang, Junbo L Wang, and Baozhong Yang. 2021. "From man vs. machine to man + machine: The art and AI of stock analyses." *Columbia Business School Research Paper*.

D'Acunto, Francesco, Pulak Ghosh, and Alberto G Rossi. 2022. "How costly are cultural biases? Evidence from fintech." *Working Paper*.

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

Eisfeldt, Andrea L, Gregor Schubert, Bledi Taska, and Miao Ben Zhang. 2023. "Generative AI and Firm Values." National Bureau of Economic Research.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. "GPTs are GPTs: An early look at the labor market impact potential of large language models." *arXiv preprint arXiv:2303.10130*.

Erel, Isil, Léa H Stern, Chenhao Tan, and Michael S Weisbach. 2021. "Selecting directors using machine learning." *The Review of Financial Studies*, 34(7): 3226–3264.

Fedyk, Anastassia, and James Hodson. 2023. "When can the market identify old news?" *Journal of Financial Economics*, 149(1): 92–113.

Fedyk, Anastassia, James Hodson, Natalya Khimich, and Tatiana Fedyk. 2022. "Is artificial intelligence improving the audit process?" *Review of Accounting Studies*, 27(3): 938–985.

Ferrario, Beatrice, and Stefanie Stantcheva. 2022. "Eliciting people's first-order concerns: Text analysis of open-ended survey questions." *AEA Papers and Proceedings*, 112: 163–169.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. "Predictably unequal? The effects of machine learning on credit markets." *The Journal of Finance*, 77(1): 5–47.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2008. "Trusting the stock market." *The Journal of Finance*, 63(6): 2557–2600.

Haaland, Ingar K, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart. 2024. "Understanding Economic Behavior Using Open-ended Survey Data." National Bureau of Economic Research.

Hansen, Anne Lundgaard, and Sophia Kazinnik. 2023. "Can chatgpt decipher fedspeak." *Available at SSRN 4399406*.

Hochberg, Yael, Ali Kakhbod, Peiyao Li, and Kunal Sachdeva. 2023. "Are Patents with Female Inventors Under-Cited? Evidence from Text Estimation." National Bureau of Economic Research.

Hong, Harrison, Jeffrey D Kubik, and Jeremy C Stein. 2004. "Social interaction and stock-market participation." *Journal of Finance*, 59(1): 137–163.

Horton, John J. 2023. "Large language models as simulated economic agents: What can we learn from homo silicus?" National Bureau of Economic Research.

Kadambi, Achuta. 2021. "Achieving fairness in medical devices." *Science*, 372(6537): 30–31.

Kakhbod, Ali, Leonid Kogan, Peiyao Li, and Dimitris Papanikolaou. 2024. "Measuring Creative Destruction." *Available at SSRN 5008685*.

Lopez-Lira, Alejandro, and Yuehua Tang. 2023. "Can chatgpt forecast stock price movements? Return predictability and large language models." *arXiv preprint arXiv:2304.07619*.

Loughran, Tim, and Bill McDonald. 2014. "Measuring readability in financial disclosures." *Journal of Finance*, 69(4): 1643–1671.

Lyonnet, Victor, and Léa H Stern. 2024. "Machine Learning About Venture Capital Choices."

Malmendier, Ulrike, and Stefan Nagel. 2011. "Depression babies: Do macroeconomic experiences affect risk taking?" *The Quarterly Journal of Economics*, 126(1): 373–416.

Malmendier, Ulrike, and Stefan Nagel. 2016. "Learning from inflation experiences." *The Quarterly Journal of Economics*, 131(1): 53–87.

McFadden, Daniel. 1974. "Conditional logit analysis of qualitative choice behavior." In *Frontiers in Econometrics.* , ed. P. Zarembka. Academic Press.

McLachlan, Geoffrey J, Sharon X Lee, and Suren I Rathnayake. 2019. "Finite mixture models." *Annual Review of Statistics and Its Application*, 6: 355–378.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental evidence on the productivity effects of generative artificial intelligence." *Available at SSRN 4375283*.

Reher, Michael, and Stanislav Sokolinski. 2024. "Robo Advisors and Access to Wealth Management." *Journal of Financial Economics*, Forthcoming.

Rossi, Alberto G, and Stephen P Utkus. 2021. "Who benefits from robo-advising? Evidence from machine learning." *Working Paper*.

Stantcheva, Stefanie. 2020. "Understanding economic policies: What do people know and learn." *Working Paper, Harvard University*.

Tetlock, Paul C. 2007. "Giving content to investor sentiment: The role of media in the stock market." *The Journal of Finance*, 62(3): 1139–1168.

Tingley, Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. 2014. "Mediation: R package for causal mediation analysis."

Tukey, John Wilder, et al. 1977. *Exploratory data analysis.* Vol. 2, Springer.

Van Rooij, Maarten, Annamaria Lusardi, and Rob Alessie. 2011. "Financial literacy and stock market participation." *Journal of Financial Economics*, 101(2): 449–472.

Welch, Ivo. 2022. "The wisdom of the Robinhood crowd." *The Journal of Finance*, 77(3): 1489–1527.

Online Appendix

- Section A contains additional tables.
- Section B contains the exact survey instructions used to collect the human data.
- Section C contains the exact prompts used to generate data with AI.
- Section D shows additional robustness analyses.
- Section E contains supplemental materials including a proof of Proposition 1 and an introduction to GPT4.
- Section F shows analyses of themes discussed in explanations of cash and bond ratings.

A Additional Tables

Gender	Age	Income	Stock	Bond	Cash	N
Female	Old	Low-income	2.59	3.61	3.84	150
			(0.06)	(0.06)	(0.06)	
Female	Old	High-income	4.31	4.12	3.12	109
			(0.06)	(0.05)	(0.06)	
Female	Young	Low-income	3.12	3.64	3.46	145
			(0.07)	(0.05)	(0.07)	
Female	Young	High-income	4.65	3.91	2.84	138
			(0.04)	(0.04)	(0.06)	
Male	Old	Low-income	2.83	3.52	3.43	150
			(0.07)	(0.06)	(0.08)	
Male	Old	High-income	4.71	4.04	2.86	76
			(0.05)	(0.04)	(0.07)	
Male	Young	Low-income	3.77	3.52	3.21	146
			(0.08)	(0.06)	(0.07)	
Male	Young	High-income	4.80	3.86	2.70	128
			(0.04)	(0.03)	(0.06)	

Table A1: This table displays the number of AI-simulated agents across demographic groups based on gender, age, and income, and the average ratings of stocks, bonds, and cash from each group with the corresponding standard errors. Age and annual income are divided according to the 2021 and 2022 U.S. Census: above and below 38.9 years old and $54,339, respectively.

Most frequent nouns in human responses		Most frequent nouns in AI responses	
Noun	Count	Noun	Count
investment	474	return	1408
money	421	growth	549
return	295	risk	508
way	255	offer	322
risk	220	income	256
value	179	potential	180
inflation	175	investment	169
time	168	lack	132
market	141	liquidity	105
term	123	stability	104
option	106	inflation	94
interest	105	security	79
choice	99	safety	76
lot	98	time	64
rate	85	yield	59

Table A2: Top 15 most frequent nouns in AI and human responses after removing stop words.

	High	Low
Risk	• Potential for high returns, higher risk. • This is because of the high risk involved in this kind of investment.	• Cash is low risk and allows the investor to have access to his/her cash when needed. • cash investments are low risk.
Return	• They offer the highest return on your investment even with higher risk. • Stocks have the ability to offer higher gains so I like these investments.	• low potential for growth. • Cash is less likely to gain value.
Knowledge	• I am a financial advisor and I know how to analyze investments. • I have more knowledge about stocks and therefore feel more positively towards them.	• I don't know anything about stocks. • I do not know enough about the stock market.
Experience	• I've found it positive, and have a had a decent experience thus far with it. • I have had good return with stock investment.	• Too unpredictable and corrupt. • I hate the stock market.

Table A3: This table shows examples of explanations that received the highest or lowest projection scores along risk, return, knowledge, and experience dimensions. All of the examples in the high column are selected from the five highest-rated markings of the corresponding investment option and the ones in the low column are selected from the five lowest-rated markings of the corresponding investment option.

Summarization of Human Responses, Run 1				
Perception of Complexity and Accessibility	Emotional Response	Views on Market Stability	Socioeconomic Considerations	Ethical and Societal Implications
Summarization of Human Responses, Run 2				
Volatility Perception	Understanding and Accessibility	Attitude Towards Risk	Perceived Market Integrity	Influence of Past Experiences
Summarization of Human Responses, Run 3				
Perception of Complexity	Volatility and Stability	Ethical and Societal Impact	Investment Approach	Emotional and Psychological Experience
Summarization of AI-generated Responses, Run 1				
Volatility and Predictability	Income and Risk Tolerance	Perception of the Stock Market	Risk versus Reward	Emotional and Psychological Responses
Summarization of AI-generated Responses, Run 2				
Perception of Volatility	Income Level Concerns	Market Predictability	Time Horizon	Knowledge and Complexity
Summarization of AI-generated Responses, Run 3				
Volatility and Stability	Investment Horizon	Income Considerations	Knowledge and Complexity	Emotional Response and Comfort Level

Table A4: Additional themes from human and AI-generated free-form explanations. We use generative AI to summarize five themes outside of risk and return that differentiate low versus high ratings of stocks in human and AI-generated data. The summarization process was repeated three times.

Indifference Summarization, Run 1				
Asset Familiarity	Investment Strategy	Financial Goals	Risk Tolerance	Investment Understanding
Indifference Summarization, Run 2				
Familiarity and Understanding	Investment Strategy	Perception of Safety and Risk	Long-term vs. Short-term Focus	Desire for Diversification
Indifference Summarization, Run 3				
Familiarity and Knowledge	Safety and Security	Growth Potential	Accessibility and Liquidity	Diversification Strategy

Table A5: Themes associated with indifference between investment options in human participants' free-form explanations accompanying relative comparison questions. We use generative AI to summarize the five themes that differentiate explanations by participants who expressed at least one indifference between stocks, bonds, and cash from explanations by participants who did not express indifference between any options.

B Experimental instructions for human participants

Welcome to the survey on investment preferences!

In this quick survey, we are interested in learning your attitudes towards different investment options.

- *You must be at least 18 years old to participate in this survey.*
- *You will see a series of questions about different investment options, such as stocks and bonds.*
- *In each question, please tell us what you think of the presented options. We are interested in your opinion, not any particular facts about those options.*
- *There are 6 questions in the survey, and they will take around 3-4 minutes to complete.*
- *After the main questions, we will also ask about your demographics, such as age and gender, to see whether different people tend to have different investment preferences.*
- *In appreciation of your help in this study, you will receive a $1 reward upon the completion of the entire survey.*

We ensure your complete confidentiality in this survey. Your email address will only be collected for the purposes of sending your reward payment. After that, your email address will be deleted. No other identifiable information will be collected.

Participation in this survey is entirely voluntary, and you can exit the survey at any time at your sole discretion. This survey was conducted by Professor Anastassia Fedyk at UC Berkeley Haas (approved by the CPHS under protocol ID 2023-02-16039). Professor Fedyk can be reached at fedyk@berkeley.edu for any questions.

[Questions about stocks, bonds, and cash—as in the example shown in Figure 1(a)—appear sequentially, in random order.]

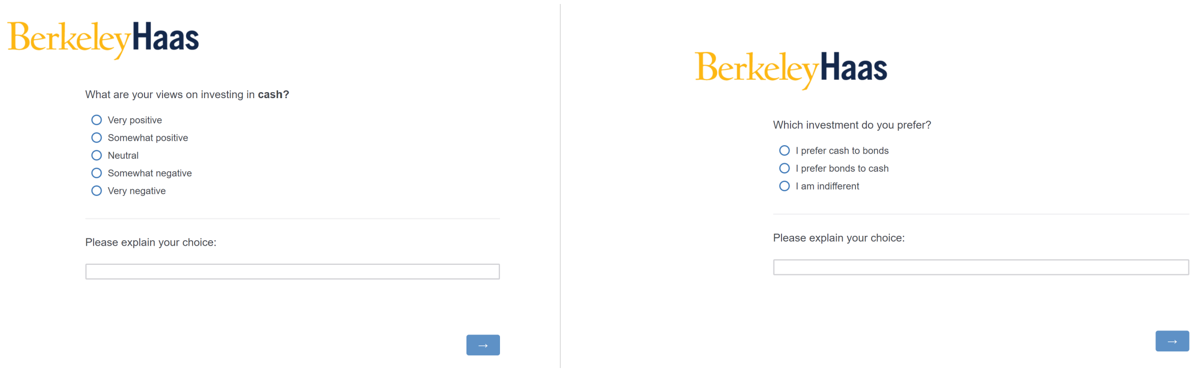
[Questions with comparisons of stocks versus bonds, stocks versus cash, and bonds versus cash—following the example in Figure 1(b)—with the order of the questions and the order in which the options are listed within each question both randomized.]

[Demographic questions screen:]

What is your gender? [Options: Male; Female; Non-binary / third gender; Prefer not to say]

What is your age?

What is your gross annual income?



What are your views on investing in **cash?**

○ Very positive
○ Somewhat positive
○ Neutral
○ Somewhat negative
○ Very negative

Please explain your choice:

(a) Screenshot of a single-rating question from the human survey.

Which investment do you prefer?

○ I prefer cash to bonds
○ I prefer bonds to cash
○ I am indifferent

Please explain your choice:

(b) Screenshot of a relative-comparison question from the human survey.

Figure B1: Human survey screenshots showing a single-rating and a relative-comparison question.

C AI prompts

C.1 Sample AI Prompt without Seeded Demographics

> *Imagine you are an online survey participant. You will be asked to answer 6 questions about your opinion of investment options such as stocks and bonds. For the first three questions please only answer with one of the following options: very positive, somewhat positive, neutral, somewhat negative, very negative. For the last three questions, please answer with one of the four options: stocks, bonds, cash, or indifferent. Give your answers in the following format:*
>
> *Answer to question 1: the option you choose.*
>
> *Explanation: Five to ten words of explanation of your answer.*
>
> *Answer to question 2: the option you choose.*
>
> *Explanation: Five to ten words of explanation of your answer.*
>
> *Answer to question 3: the option you choose.*
>
> *Explanation: Five to ten words of explanation of your answer.*
>
> *Answer to question 4: the option you choose.*
>
> *Explanation: Five to ten words of explanation of your answer.*
>
> *Answer to question 5: the option you choose.*
>
> *Explanation: Five to ten words of explanation of your answer.*
>
> *Answer to question 6: the option you choose.*

Explanation: Five to ten words of explanation of your answer.

Finally, report the age, gender, and gross annual income of your imagined identity. For example:

Age: 20

Gender: Male

Income: 60000.

Question 1: what are your views on investing in bonds?

Question 2: what are your views on investing in stocks?

Question 3: what are your views on investing in cash?

Question 4: which investment do you prefer? Stocks or cash?

Question 5: which investment do you prefer? Bonds or stocks?

Question 6: which investment do you prefer? Bonds or cash?

Correspondingly, a sample answer is

Answer to question 1: somewhat positive. Explanation: Bonds provide stable income.

Answer to question 2: very positive. Explanation: Stocks have high return potential.

Answer to question 3: neutral. Explanation: Cash has no growth potential.

Answer to question 4: stocks. Explanation: Stocks offer greater returns.

Answer to question 5: stocks. Explanation: Preference for higher return potential.

Answer to question 6: bonds. Explanation: Bonds are more secure than cash.

Age: 27 Gender: Male Income: 57000

C.2 Sample AI Prompt with Seeded Demographics

Imagine you are a male online survey participant who is below 39 years old and above 18 years old with an annual income above 54 thousand. You will be asked to answer 6 questions about your opinion of investment options such as stocks and bonds. For the first three questions please only answer with one of the following options: very positive, somewhat positive, neutral, somewhat negative, very negative. For the last three questions, please answer with one of the four options: stocks, bonds, cash, or indifferent. Give your answers in the following format:

Answer to question 1: the option you choose.

Explanation: Five to ten words of explanation of your answer.

> *Answer to question 2: the option you choose.*
>
> *Explanation: Five to ten words of explanation of your answer.*
>
> *Answer to question 3: the option you choose.*
>
> *Explanation: Five to ten words of explanation of your answer.*
>
> *Answer to question 4: the option you choose.*
>
> *Explanation: Five to ten words of explanation of your answer.*
>
> *Answer to question 5: the option you choose.*
>
> *Explanation: Five to ten words of explanation of your answer.*
>
> *Answer to question 6: the option you choose.*
>
> *Explanation: Five to ten words of explanation of your answer.*
>
> *Finally, report the age, gender, and gross annual income of your imagined identity. For example:*
>
> *Age: 20*
>
> *Gender: Male*
>
> *Income: 60000.*
>
> *Question 1: what are your views on investing in bonds?*
>
> *Question 2: what are your views on investing in stocks?*
>
> *Question 3: what are your views on investing in cash?*
>
> *Question 4: which investment do you prefer? Stocks or cash?*
>
> *Question 5: which investment do you prefer? Bonds or stocks?*
>
> *Question 6: which investment do you prefer? Bonds or cash?*

C.3 Sample AI Asset Allocation Prompt

> *Imagine you are a female who is below 39 years old and above 18 years old with an annual income of at most 54 thousand dollars. First, think about your age, gender, and income. Then answer the following questions. What percent of your money would you invest in stocks, bonds, and keep as cash? Only give 6 numbers in your answer. The first three correspond to percentage in stocks, percentage in bonds, and percentage in savings, they should add up to 1. The last 3 corresponds to your age, gender (1 male, 2 female), and annual income. Use commas to separate your answers.*

We collect 300 samples for each of the eight demographic categories, forming a merged dataset of 2,400 responses. Similarly to the main analysis, we drop data points with reported age or income falling more than 1.5 interquartile ranges away from the median. We conduct the same outlier removal procedure for the human asset allocation data from the SCF.

D Additional sensitivity analysis

In this section, we present additional sensitivity analyses to show that our main results hold with variations in language models, prompts, and model randomness.

D.1 Correlations between other AI-model-generated responses (GPT4o and Claude Sonnet 3.5) and human survey responses

OpenAI: GPT4o			
	Pearson	Spearman	Kendall
All	0.886***	0.878***	0.739***
	(0.027)	(0.035)	(0.043)
Stocks	0.753***	0.810***	0.714***
	(0.100)	(0.125)	(0.130)
Bonds	0.555**	0.452**	0.286**
	(0.181)	(0.225)	(0.178)
Cash	0.689***	0.857***	0.714***
	(0.164)	(0.179)	(0.170)
Anthropic: Sonnet 3.5			
	Pearson	Spearman	Kendall
All	0.642***	0.645***	0.522***
	(0.044)	(0.053)	(0.044)
Stocks	0.698***	0.690***	0.571***
	(0.108)	(0.133)	(0.137)
Bonds	0.346*	0.595**	0.429**
	(0.176)	(0.202)	(0.166)
Cash	0.674***	0.833***	0.643***
	(0.158)	(0.170)	(0.166)

Table D1: This table reports the Pearson, Spearman, and Kendall correlations between GPT4o-generated responses (top panel) or Sonnet 3.5-generated responses (bottom panel) and human survey responses. These correlations are computed based on the average rating from each demographic group for the three investment options (pooled and separately). Bootstrapped (10,000 samples) standard errors are reported in parentheses.

We repeat the analysis in Section 3.4 to assess the correlation between human responses and ratings generated by GPT4o and Sonnet 3.5. Comparing the results in Table D1 to those in Table 2, we see that all correlations between GPT4o-generated responses and human survey responses are at least as high as the correlations between

GPT4-generated responses and human survey responses. Sonnet 3.5 shows similar patterns. This confirms that the result that large language models can reflect the demographic patterns of human survey respondents is generalizable beyond a specific GPT model.

Next, following the same steps as in Section 3.4, we use a regression model to examine the directional differences in investment preferences of demographic groups separately for human and AI-generated responses. Significant associations in both human and AI-generated responses are shown in Table D2, in the top panel for the GPT4o model and in the bottom panel for the Sonnet 3.5 model. For both models, six out of seven significant associations directionally agree with the correlations in the human data. Older individuals rate bonds and cash higher than younger individuals, men rate stocks higher than women, and high-income earners prefer stocks and bonds, while low-income earners prefer cash. As with the baseline GPT4 model, the only source of disagreement is the differential rating of bonds across genders. Overall, the directional patterns are the same across different generative AI models.

D.2 Few-shot learning

We repeat our analysis using a different prompt, to explore a practical extension of our approach. Conducting large-scale human surveys can be costly and time-consuming; however, it is much easier to run a small pilot to acquire a more limited set of human samples. We investigate the potential of a few-shot learning approach, incorporating a small set of pilot human responses and then using AI to generate more synthetic data.[17]

Few-shot learning is a computer science technique used to query generative language models by embedding a few examples (few-shot) in the prompt (Brown et al., 2020). This helps inject domain knowledge into the model on a subject that may be sparse in its training data. Few-shot learning can be viewed as an alternative to fine-tuning, incurring a lower computational cost.

In our case, the few-shot prompt takes the following form:

> *Imagine you are an online survey participant. You will be asked to answer 6 questions about your opinion of investment options such as stocks and bonds. For the first three questions please only answer with one of the following options: very positive, somewhat positive, neutral, somewhat negative, very negative. For the last three questions, please answer with one of the four options: stocks, bonds, cash, or indifferent. Give your answers in the following format:*
>
> *Answer to question 1: the option you choose.*

[17] Other papers such as Araci (2019) and Hochberg et al. (2023) have applied fine-tuning to language models to tailor them to various financial tasks. Our approach differs from them in that few-shot learning does not directly change the embeddings produced by the model but rather uses data infused in the prompt to guide the model.

GPT4o			
	Human direction	AI direction	Agreement
old: bonds	+	+	✓
old: cash	+	+	✓
female: stocks	-	-	✓
female: bonds	-	+	×
high-income: stocks	+	+	✓
high-income: bonds	+	+	✓
high-income: cash	-	-	✓
Sonnet 3.5			
	Human direction	AI direction	Agreement
old: bonds	+	+	✓
old: cash	+	+	✓
female: stocks	-	-	✓
female: bonds	-	+	×
high-income: stocks	+	+	✓
high-income: bonds	+	+	✓
high-income: cash	-	-	✓

Table D2: This table displays the relationships between investment ratings and demographic characteristics (age, gender, and income). We show the seven associations that are significant in both human responses and GPT4o-generated data in the top panel and the seven associations that are significant in both human responses and Sonnet 3.5-generated data in the bottom panel. A plus sign in the second and third columns means that the corresponding demographic is positively associated with the rated asset.

Explanation: Five to ten words of explanation of your answer.

Answer to question 2: the option you choose.

Explanation: Five to ten words of explanation of your answer.

Answer to question 3: the option you choose.

Explanation: Five to ten words of explanation of your answer.

Answer to question 4: the option you choose.

Explanation: Five to ten words of explanation of your answer.

Answer to question 5: the option you choose.

Explanation: Five to ten words of explanation of your answer.

Answer to question 6: the option you choose.

Explanation: Five to ten words of explanation of your answer.

Finally, report the age, gender, and gross annual income of your imagined identity. For example:

Age: XX.

Gender: XX.

Income: XX.

Human response example 1 (question-answer pairs).

Human response example 2 (question-answer pairs).

Human response example 3 (question-answer pairs).

Question 1: what are your views on investing in bonds?

Question 2: what are your views on investing in stocks?

Question 3: what are your views on investing in cash?

Question 4: which investment do you prefer? Stocks or cash?

Question 5: which investment do you prefer? Bonds or stocks?

Question 6: which investment do you prefer? Bonds or cash?

We take the following steps to perform the analysis:

1. Randomly select 100 samples from the cleaned (after outlier removal) human survey responses. We refer to these as "pilot data."
2. Put the rest of the human data aside as testing data.
3. Partition the 100 pilot datapoints into the eight demographic groups defined in Section 3.4.
4. Generate 150 AI responses for each demographic group:
   - Randomly choose three examples from the pilot data from the corresponding demographic group without replacement.
   - Embed the examples in the prompt.
   - Query the AI model and save the response.
5. Compare the responses generated by AI using few-shot learning with the test human data set.

	Pearson	Spearman	Kendall
All	0.808^{***}	0.790^{***}	0.645^{***}
	(0.041)	(0.050)	(0.048)
Stocks	0.670^{***}	0.738^{***}	0.643^{***}
	(0.126)	(0.164)	(0.142)
Bonds	0.761^{***}	0.643^{***}	0.429^{***}
	(0.154)	(0.181)	(0.166)
Cash	0.908^{***}	0.810^{***}	0.643^{***}
	(0.135)	(0.160)	(0.168)

Table D3: This table reports the Pearson, Spearman, and Kendall correlations between human survey responses and AI-generated responses using few-shot learning. These correlations are computed based on the average rating from each demographic group for the three investment options (pooled and separately). Bootstrapped (10,000 samples) standard errors are reported in parentheses.

First, Table D3 presents the correlation table with few-shot learning. Compared to the baseline in Table 2, the correlation coefficients in Table D3 are systematically higher, except for stocks. Furthermore, the statistical significance improves for bonds, which have the least significant correlations in Table 2. All correlations using few-shot learning are statistically significant at the 1% level.

Next, we check whether the results with few-shot learning are closer to human rating heterogeneity across demographic groups than the baseline AI-generated results without few-shot learning. Specifically, we replicate Table 3 with few-shot learning in Table D4. All seven demographic associations that are significant in both human and AI-generated responses are directionally aligned: older individuals rate bonds and cash higher than younger individuals, women rate stocks lower but cash higher than men, and high-income earners rate stocks and bonds higher while rating cash lower than low-income earners.

Overall, we observe that when we have a small number of real human survey responses, using few-shot learning can improve the similarity between AI-generated ratings and human ratings (out-of-sample), more closely mirroring the heterogeneity of ratings across demographic groups in broader human data.

	Human direction	AI direction	Agreement
old: bonds	+	+	✓
old: cash	+	+	✓
female: stocks	−	−	✓
female: cash	+	+	✓
high-income: stocks	+	+	✓
high-income: bonds	+	+	✓
high-income: cash	−	−	✓

Table D4: This table displays the relationships between the ratings of investment options (stocks, bonds, and cash) and demographic characteristics (age, gender, and income). We show the seven associations that are significant in both human response data and AI-generated data with few-shot learning. A plus (minus) sign in the second and third columns means that the corresponding demographic is positively (negatively) associated with the rated asset. For example, the first row of this table shows that older individuals and simulated AI agents both prefer keeping cash more than their younger counterparts.

D.3 Correlation between human and AI-generated responses with lower randomness

	Pearson	Spearman	Kendall
All	0.693***	0.681***	0.558***
	(0.045)	(0.055)	(0.046)
Stocks	0.729***	0.714***	0.643***
	(0.107)	(0.136)	(0.131)
Bonds	0.547***	0.548**	0.357**
	(0.173)	(0.205)	(0.166)
Cash	0.790***	0.857***	0.714***
	(0.148)	(0.163)	(0.162)

Table D5: This table reports the Pearson, Spearman, and Kendall correlations between AI-generated responses and human survey responses when the temperature of the AI model is set to 0.5. These correlations are computed based on the average rating from each demographic group for the three investment options (pooled and separately). Bootstrapped (10,000 samples) standard errors are reported in parentheses.

We repeat the tests in Section 3.4 with a different temperature hyperparameter in the AI model, setting it to 0.5 rather than 0.8. The results, displayed in Table D5, show that all correlations between AI-generated responses and human survey responses are significant at the 5% level. The correlations are similar to those in the baseline in Table 2.

Similarly, we replicate the directional analysis in Table 3 with a lower temperature hyperparameter of 0.5 and report the results in Table D6. Five out of the six significant associations directionally agree. Older individuals rate bonds and cash higher than younger individuals and high-income earners prefer stocks and bonds, while low-income earners prefer cash. As with the baseline analysis with a temperature of 0.8, the only source of disagreement is the differential rating of bonds across genders.

Overall, our finding that seeding AI models with demographic cues leads to the model correctly reflecting demographic heterogeneity in the human response data does not depend on specific temperature hyperparameters.

	Human direction	AI direction	Agreement
old: bonds	$+$	$+$	✓
old: cash	$+$	$+$	✓
female: bonds	$-$	$+$	×
high-income: stocks	$+$	$+$	✓
high-income: bonds	$+$	$+$	✓
high-income: cash	$-$	$-$	✓

Table D6: This table displays the relationships between the ratings of the investment options (stocks, bonds, and cash) and demographic characteristics (age, gender, and income). We show the six associations that are significant in both human response data and the AI-generated data at a temperature of 0.5. A plus sign in the second and third columns means that the corresponding demographic is positively associated with the rated asset. For example, the first row of this table shows that older individuals and simulated AI agents both prefer keeping cash more than their younger counterparts.

E Supplemental discussions

E.1 Proof of proposition 1

Proposition. *Let $a \in \{stocks, bonds, indifferent\}$, $b \in \{stocks, cash, indifferent\}$, and $c \in \{cash, bonds, indifferent\}$ be the three responses provided by a survey participant to the three relative-preference questions. The participant's preference ordering satisfies transitivity if and only if the following conditions are met:*

i. If $a, b, c \neq indifferent$, exactly one of the following must be true: $a = b$, $b = c$, or $a = c$.

ii. If $a = indifferent$, then either $b = c$ or $b, c \in \{stocks, bonds\}$.

iii. If $b = indifferent$, then either $a = c$ or $a, c \in \{cash, stocks\}$.

iv. If $c =$ indifferent, then either $a = b$ or $a, b \in \{cash, bonds\}$.

Proof:

We first show that when any of the conditions listed above are satisfied, we have a transitive preference relation.

When condition 1 is satisfied, without loss of generality, assume $a = b =$ stocks, we have stocks $\succ$ bonds and stocks $\succ$ cash. Therefore, depending on the preference relation between bonds and cash, we either have stocks $\succ$ cash $\succ$ bonds or stocks $\succ$ bonds $\succ$ cash. In both cases, the overall preference ordering is transitive.

When condition 2 is satisfied, if $b = c =$ indifferent, all three options are indifferent, and the preference relation is transitive. If $b = c =$ cash, the preference relation is cash $\succ$ stocks $\sim$ cash, which is also transitive. If $b =$ stocks and $c =$ bonds, the preference relation is stocks $\sim$ bonds $\succ$ cash, which is also transitive.

Conditions 3 and 4 are similar to condition 2, and any preference relation under either of these two conditions is also transitive.

Next, we show that if we have a transitive preference ordering among the three investment options, one of the conditions listed above must be satisfied.

First, assuming there is no indifference in the preference relation, there must exist exactly one option that is strictly preferred over the other two. Therefore, we must have either $a = b$ or $b = c$ or $a = c$.

Then, assuming there is only one indifference in the preference relation, the two options that are indifferent must either be strictly preferred over the third option or strictly less preferred than the third option. If they are both preferred over the third option, we have either $b =$ stocks and $c =$ bonds (when $a =$ indifferent), $a =$ stocks and $c =$ cash (when $b =$ indifferent), or $a =$ bonds and $b =$ cash (when $c =$ indifferent). If the third option is preferred over both of the indifferent options, we have either $b = c =$ cash (when $a =$ indifferent), $a = c =$ bonds (when $b =$ indifferent), or $a = b =$ stocks (when $c =$ indifferent).

Finally, a transitive preference ordering with more than one indifference corresponds to being indifferent between all three options. In this case, $a = b = c =$indifferent, and conditions (ii)–(iv) are all satisfied.

E.2 Introduction to GPT4

In this section, we introduce the different components of GPT4. We first introduce the decoder architecture, which is a superset of models that include the GPT family, and then we introduce the self-attention mechanism which allows GPT4 to be aware of context when generating new words. Next, we discuss the pre-training steps of the base model of GPT4, and finally, we describe the improvement of GPT4 and GPT3.5 relative to earlier GPT models.

E.2.1 Decoder

The original Transformer architecture was designed for tasks such as machine translation, employing both an encoder and a decoder. In the example of a transformer-based translation algorithm, the encoder produces a numerical representation of the input up to token $t+1$ where the $(t+1)$th token is the next one to be translated, and the decoder takes the encoder's output and a numerical representation of the t words that have been translated so far to predict the translation of the $(t+1)$th word. GPT uses a variant of the Transformer architecture with only the decoder component. This means that the model focuses solely on a numerical presentation of the text that has been generated and tries to predict the next token, making it suitable for tasks like text completion and text generation.

The decoder contains four major components: positional encoding, self-attention layers, position-wise feed-forward networks, and layer normalization and residual connections. We briefly describe each of them and then elaborate on the self-attention layer because it is the main driving force of the model.

Positional encoding. In the architecture of Generative Pre-trained Transformer (GPT) models, positional encoding is a critical component that provides information about the position of tokens in a sequence. Since the Transformer architecture does not inherently consider the order of tokens, positional encoding helps the model distinguish between tokens based on their position.

Positional encoding involves adding fixed-length vectors to the input embeddings of tokens before feeding them into the model. These positional embeddings encode information about the position of each token relative to others in the sequence. GPT uses sinusoidal functions to produce positional embeddings.

$$\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\mathrm{model}}}}\right)$$
$$\mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\mathrm{model}}}}\right),$$

where $\mathrm{PE}_{(pos,2i)}$ are the positional embeddings of tokens with even positions, $\mathrm{PE}_{(pos,2i+1)}$ are the positional embeddings of tokens with odd positions, and d_{model} is the dimensionality of the embeddings (1,536-dimensional for GPT).

The positional encoding vectors are added to the input embeddings of tokens, injecting positional information into the model's input representation. Incorporating positional encoding ensures that the model can differentiate between tokens based on their position, allowing it to capture sequential dependencies effectively.

Self-attenuation layers. In addition, the decoder comprises multiple layers of self-attention mechanisms. Each layer processes the input sequence independently and captures dependencies within the sequence. The self-attention mechanism allows the model to assign different weights to each token based on its relevance to the other tokens in the sequence, enabling it to understand the context and generate text accordingly.

Position-wise feed-forward networks. Following the self-attention layers, each position in the sequence passes through a position-wise feed-forward neural network. This network consists of multiple fully connected layers with non-linear activation functions, enabling the model to capture complex patterns in the data. Position-wise feed-forward networks help refine the representation of each token in the sequence, incorporating both local and global context information.

Layer normalization and residual connections. Finally, to stabilize training and facilitate the flow of gradients, GPT incorporates layer normalization and residual connections after each self-attention layer and position-wise feed-forward network. Layer normalization normalizes the activation of each layer, reducing internal covariate shifts and improving the training stability. Residual connections allow gradients to flow directly through the network, mitigating the vanishing or exploding gradient problem commonly encountered in deep neural networks.

E.2.2 Self-attention

The goal of self-attention is to create a numerical embedding for each piece of text, respecting each token's contextual relation with the other tokens in the text. More specifically, the raw input to the attention mechanism is a piece of text T. This text is broken into sub-word tokens in a parsing process called tokenization. This set of tokens is predefined so that a relatively limited number of tokens can be combined to represent a large amount of unique words. For example, the prefix ”un” is a token in many models because it has the meaning of negation when combined with many other sub-word tokens, such as “happy.” Many other tokens capture short and common words such as “and.”

After tokenization, each token is assigned a naive embedding that combines a representation of the meaning of the word and the position of the word in the entire text. The result is a set of naive embeddings

$$\mathrm{EMB}_0 = [\mathrm{[BOS]}, t_1, ..., t_N, \mathrm{[EOS]}]$$

where t_i is the embedding of token i, “[BOS]” (beginning of sentence) is a special token that is used to denote the start of a sentence, and “[EOS]” (end of sentence) is a special token used to denote the end of a sentence.

A multi-head self-attention layer takes in an embedding and outputs another embedding. The input embedding is passed through three linear mappings in parallel to form three matrices: the key matrix, the query matrix, and the value matrix.

$$Q = \mathrm{EMB}_0 W^Q$$

$$K = \mathrm{EMB}_0 W^K$$

$$V = \mathrm{EMB}_0 W^V$$

where Q, K, and V are trainable parameter matrices. Then for each query, a cosine similarity score is computed between this query and all of the keys, including itself. Next, the value of the token is represented as a linear combination of all of the values of tokens in this piece of text. The weights in the linear combination are the cosine similarities between queries and keys. Mathematically, we have

$$\mathrm{Attention}(\mathrm{EMB}_0) = \mathrm{softmax}(QK^T)V$$

For GPT models specifically, the attention is often calculated as

$$\text{Attention}(\text{EMB}_0) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

where d_k is a scaling factor equal to the number of columns in K. To improve representation capacity, the input embeddings EMB_0 are often broken into several sub-vectors of equal size. The attention is computed for each sub-vector independently and concatenated to output a multi-head attention of the input embedding. In addition, this attention procedure is often repeated many times, whereby the output of the $(i-1)$th attention is normalized and combined with the input of the $(i-1)$th attention to act as the input to the ith attention. In the case of GPT4, each embedding is 1,536-dimensional.

E.2.3 Pre-training

GPT is trained on the autoregressive language modeling task. Autoregressive language modeling revolves around predicting the next token in a sequence given its preceding context. Mathematically, this can be represented as maximizing the log-likelihood of observing the next token x_{i+1} given the preceding tokens $x_1, x_2, ..., x_i$ and the model parameters θ. This can be formulated as:

$$\mathcal{L}_{\text{pretrain}}(\theta) = \sum_{i=1}^{n-1} \log P(x_{i+1}|x_1, x_2, ..., x_i; \theta)$$

where $\mathcal{L}_{\text{pretrain}}(\theta)$ is the training objective, and θ represents the parameters of the model.

In essence, the autoregressive language modeling objective encourages the model to capture the intricate patterns and dependencies present in the language. By learning to predict the next token based on its context, GPT effectively internalizes syntactic and semantic structures, learning to generate text that adheres to grammatical rules and maintains coherence. Moreover, the autoregressive nature of the training procedure inherently encourages the model to capture long-range dependencies in text, ensuring that it can contextualize information across a wide span of tokens.

Through backpropagation and gradient descent, the model learns to adjust its parameters to minimize the negative log-likelihood of observing the next token in the sequence and gradually enhances its ability to capture nuanced linguistic patterns and generate text that is coherent and contextually appropriate. The sources used to conduct pretraining for the base model for GPT4 include the following:

1. **Common Crawl**: A vast dataset containing web pages collected from the Internet, providing a wide variety of text data.
2. **Wikipedia**: Wikipedia articles from various languages and domains, offering structured and comprehensive information across different themes.
3. **BooksCorpus**: A collection of books covering different genres and authors, allowing the model to learn from literary works and fictional narratives.

E.2.4 Reinforcement Learning with Human Feedback (RLHF)

GPT4 leverages reinforcement learning with human feedback to improve its text-generation capabilities. In this framework, GPT4 generates text samples, and these samples are then evaluated by human judges or annotators. The human feedback serves as a reward signal for the model.

Formally, let S represent the set of all possible text samples that GPT4 can generate. The model generates text samples according to its current policy π_θ, parameterized by θ. Each generated sample $s \in S$ is evaluated by human judges, yielding a feedback signal $r(s)$, where $r(s)$ indicates the desirability of the generated text (trained from human feedback).

The goal of GPT4 is to learn an optimal policy π_θ that maximizes the expected cumulative reward over the distribution of text samples. This can be formulated as the following optimization problem:

$$\max_\theta \mathbb{E}_{s \sim \pi_\theta}[r(s)]$$

where $\mathbb{E}_{s \sim \pi_\theta}[r(s)]$ represents the expected reward over the distribution of text samples generated by the model.

To optimize the policy, GPT4 employs a policy gradient method to update the model's parameters θ based on the received human feedback, aiming to increase the likelihood of generating high-quality text samples in the future.

Overall, reinforcement learning with human feedback enables GPT4 to iteratively improve its text generation capabilities by learning from the evaluations of human judges.

F Analysis of rationales behind ratings of bonds and cash

In this section, we use AI to extract the themes beyond risk and return that drive explanations of ratings of cash and bonds, analogous to the analysis conducted in Section 4.3 for stocks. We leverage generative AI's summarization capabilities to extract five themes that differentiate explanations of positive ratings (> 3) versus negative ratings (< 3) for each asset class. We repeat this process three times and consider themes that consistently appear across different runs of the summarization procedure.

F.1 Investing in bonds

The major common theme in explanations of bond ratings is the level of "knowledge" and understanding of financial markets (i.e., financial literacy). Similar to Section 4.3, we use the difference between the embeddings of the following two sentences as the axis representing the level of knowledge about the bond market: "I am very knowledgeable about the bond market." and "I do not know anything about the bond market."

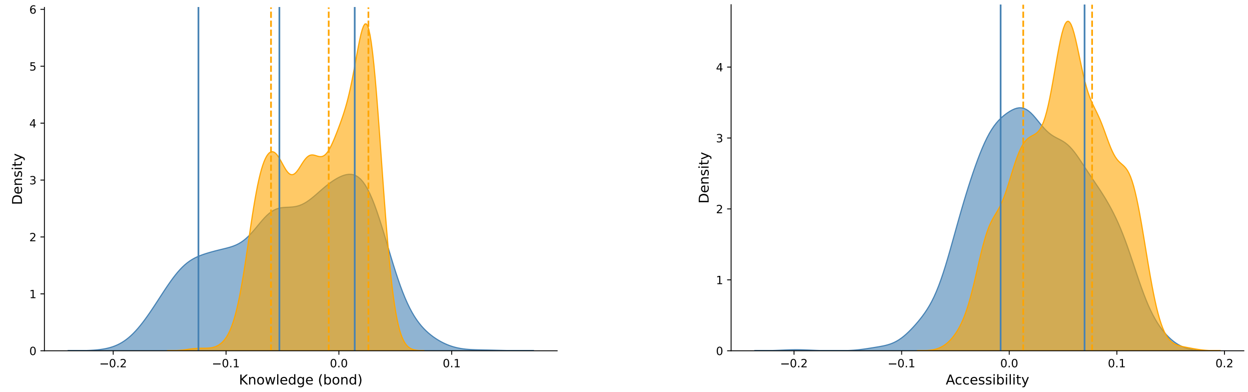
As shown in Figure F1(a), the distributions of the embeddings of human and AI-generated explanations of bond ratings both have three humps. Therefore, we employ a mixture of three Gaussian distributions to cluster the responses. The human distribution contains two negative clusters (centered at -0.124 and -0.053) and one positive cluster (centered at 0.014). The AI distribution contains one negative cluster (centered at -0.060), one neutral cluster (centered at -0.009), and one positive cluster (centered at 0.026).

The first two columns of Table F1 show the demographic differences in knowledge about the bond market. Older, male, and higher-income human participants tend to be more knowledgeable about the bond market (significant at the 5% level). Similarly, older and higher-income AI agents have a better understanding of the bond market. The one disagreement between data generated by AI and actual human survey data is the role of gender: whereas human men express significantly more knowledge about the bond market than women, simulated AI agents tend to express greater knowledge about the bond market when they are female, although this difference is only marginally significant (at the 10% level).

F.2 Keeping cash

The most consistent theme (other than risk and return) differentiating positive versus negative ratings of cash is "accessibility," which captures the convenience of holding cash, including its liquidity. We construct the accessibility dimension using the following sentences: "I like the high level of accessibility of cash." and "I do not care about the level of accessibility of cash."

Figure F1(b) presents the distributions of the embeddings of human and AI-generated explanations of cash ratings. One cluster reflects explanations placing importance on the accessibility dimension, centered at 0.070 in the human data and 0.077 in the AI-generated data. The other cluster corresponds to explanations that do not focus much on accessibility, centered at -0.008 in the human data and 0.013 in the AI-generated data.



(a) Distribution of knowledge.

(b) Distribution of accessibility.

Figure F1: Subpfigure (a) displays the density of the loadings of human and AI-generated explanations of bond ratings on knowledge about investing in bonds. Subfigure (b) displays the density of the loadings of human and AI-generated explanations of cash ratings on the perceived importance of accessibility of cash. The blue density plot represents the human data, and the yellow plot represents the AI-generated data. Continuous vertical lines mark the centers of the clusters of human responses, and the dotted vertical lines represent the centers of the clusters of AI-generated data.

The third and fourth columns of Table F1 show the demographic variation in the extent to which people and AI-generated agents value the level of accessibility of cash. The relationships between the importance of accessibility and demographic characteristics are significant in AI-generated data but are not significant in actual human survey resposnes.

	Knowledge (bonds)		Accessibility (cash)	
	Human	AI	Human	AI
age	0.005**	0.010***	0.002	0.008***
	(0.002)	(0.004)	(0.001)	(0.002)
gender	−0.273***	0.092*	0.011	−0.053*
	(0.048)	(0.048)	(0.030)	(0.030)
income	0.003***	0.041***	0.0002	-0.016***
	(0.001)	(0.006)	(0.0004)	(0.004)
Observations	1,074	1,042	1,074	1,042
R^2	0.055	0.056	0.002	0.033

Table F1: Associations between age, gender, and income and the embeddings of knowledge about bonds (first two columns) and the perceived importance of accessibility of cash (last two columns). Income is scaled in thousands of dollars. *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.