ChatGPT 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

Does generative AI accurately capture demographic heterogeneity in investment preferences? We survey over 1,200 online participants and compare human and AI responses side by side, examining ratings, free-form explanations, and pairwise comparisons across stocks, bonds, and cash. Our results indicate that default AI agents skew toward young, high-income profiles. However, when provided with demographic cues, AI reflects the same heterogeneities observed in human investment preferences—both in numerical ratings and in the rationale embedded in textual explanations. Free-form explanations also shed light on a key difference between human and AI-generated responses: humans are more prone to violating the transitivity axiom. We show that this stems from limited financial knowledge and is mediated by expressed indifference between at least two investment options.

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

JEL Classification: C1, G10, G11, G12.

*This Draft: February 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). We are grateful to Elena Asparouhova, Francesco D'Acunto, Amir Kermani, Dmitry Livdan, Alejandro Lopez-Lira, Alberto Rossi, Michael Reher, Gregor Schubert, Stas Sokolinski, and participants at NYU, Stanford University, UC Berkeley, and Carey Finance Conference for very helpful comments.

1 Introduction

The financial services industry is undergoing rapid automation, with robo-advising and online trading platforms such as Robinhood significantly increasing retail investor participation. Over the past decade, automated investment advice has grown more than tenfold, and recent advances in artificial intelligence (AI), particularly large language models, promise to further reshape the financial sector.[1] AI-driven systems have the potential to expand access to low-cost automated investment products, but they also raise concerns about algorithmic bias, particularly if training data do not adequately represent diverse investor populations. Concerns about algorithmic bias are especially relevant in investing, which features a large imbalance in demographic characteristics such as gender (Barber and Odean, 2001). To date, the introduction of machine learning has been found to disproportionately benefit white borrowers in credit screening applications (Bartlett et al., 2022; Fuster et al., 2022), and existing robo-advising tools show uneven gains across age groups (Reher and Sokolinski, 2024).

In this paper, we evaluate the extent to which state-of-the-art generative AI models can replicate human investment preferences. We conducted a large-scale survey with 1,272 human participants and 1,350 generated AI agents to the same questions, both with and without demographic information. Without demographic cues, 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. Notably, AI not only reproduces human response patterns but also captures how free-form discussions of risk and return influence investment ratings and how financial knowledge and experience vary across demographics.

Furthermore, our analysis of direct-comparison responses reveals a key distinction between human and AI-generated preferences: AI exhibits transitive orderings 98.7% of the time, compared to 84.4% for humans. Among human respondents, intransitive preferences are most prevalent among those expressing indifference between investment options. Textual analysis of free-form responses shows that violations of transitivity among human survey participants are associated with a lack of financial knowledge mediated by expressed indifference between at least two investment options.

[1] For the effects of AI in finance see, among others, Abis and Veldkamp (2024); Bybee (2023); Cao et al. (2021); Lopez-Lira and Tang (2023). In addition, Fedyk et al. (2022) show that AI increases quality and efficiency in financial audits, and Cao et al. (2021) document improved performance among financial analysts.

Our human survey was administered in two waves, October 2023 and March 2024, and consisted of three components: (i) categorical ratings of three investment options—stocks, bonds, and cash—-on a scale from "very negative" (encoded as 1) to "very positive" (encoded as 5); (ii) a free-form written explanation of each rating; and (iii) a comparison of each pair of investment options (stocks versus bonds, stocks versus cash, and bonds versus cash), choosing either the first option, the second option, or "indifferent." The respondents of the human survey (and the AI agents) are balanced across the key demographics, with 49.8% (50%) identifying as male, a median age of 37 (39), and a median personal income of $53,000 ($55,000). We queried OpenAI's GPT4 model using the same survey questions and demographics, with 1,200 simulated runs. That is, we seeded the prompts with different demographic characteristics and described the type of questions and permitted answers (rating versus free-form). We then asked the exact same questions as in the human survey. We also ran additional queries (150 simulated runs) that did *not* include demographic information.

We compare the human survey responses and the AI-generated responses along three dimensions. First, we analyze the ratings each human respondent or simulated AI agent assigns to each investment option, examining how the human ratings vary across demographics and the extent to which these patterns are captured by the AI-generated responses. Second, we perform a textual analysis of the rationale (free-form explanations) behind the ratings to study whether AI is able to capture the main themes in the justifications that human participants provide for their investment choices. Third, we use the relative comparison questions to examine the transitivity of the preference orderings in the human versus AI-generated survey data, pointing to a potential advantage of using AI-enhanced tools. Moreover, we document that the lack of financial knowledge, mediated by their indifference between investment options, is a key reason for intransitive preferences in human responses.

For the first step, the comparison of investment ratings by human participants versus AI, we form eight demographic groups categorized by gender (men or women), age (above or below the median), and income (above or below the median). For example, if a male participant has above-median income and is above-median age, we categorize him as an older, high-income (wealthy) man.

We begin by analyzing how well generative AI matches the heterogeneity of human preferences *in absence* of explicitly specifying the demographics in the AI model—that is, using the responses of 150 default AI agents. The self-reported demographics of genera-

tive AI in this sample are 56% female, 94% above-median age, and all with above-median income. In terms of the asset ratings, we form eight three-dimensional Gaussian distributions (with the three dimensions being the ratings on stocks, bonds, and cash)—one distribution for each human demographic group (e.g., young high-income women, old high-income men, etc.). For each of the 150 simulated AI agents, we identify the human distribution from which the corresponding responses would be most likely to be drawn. The default AI-generated responses overwhelmingly (in 83% of all cases) resemble young high-income individuals (a demographic group that comprises less than 30% of the population). Thus, absent a specification of demographic characteristics of the target population, generative AI models display algorithmic bias in the form of mostly reflecting the investment preferences of the overrepresented group—in this case, young high-income people.

Once the model is provided with demographic information, the algorithmic bias disappears, allowing AI to accurately replicate differences across demographic groups. To demonstrate this, we analyze a larger dataset of 1,200 AI-generated responses, with 150 responses for each of the eight demographic groups. We construct two vectors: one representing the average ratings given by human survey participants and the other representing AI-generated ratings. Each vector contains 24 numerical values, corresponding to the average ratings of the three investment options across the eight demographic groups. The results show a strong alignment between AI and human responses, with a Pearson correlation of 0.73 and a Spearman correlation of 0.70. This close match is especially remarkable given the skew in the default model without demographic information. For example, simulated generative AI agents in the default model practically never self-identify as having below-median income. And yet, when prompted to assume the identity of a particular demographic group, including low-income individuals, generative AI does a remarkable job in matching that group's preferences.

Zooming in on specific differences in the ratings across demographic groups, we confirm that male human survey participants rank stocks higher than female participants, and the reverse for bonds, consistent with prior evidence (Agnew, Balduzzi and Sunden, 2003). Higher-income individuals rate stocks and bonds higher and rank cash lower than lower-income individuals (Hong, Kubik and Stein, 2004). Finally, older individuals rate cash higher than younger individuals. Almost all of these patterns are correctly captured by AI-generated responses, which predict higher stock ratings for men, higher stock and bond ratings for higher-income individuals, lower cash ratings for higher-income indi-

viduals, and higher cash ratings for older individuals. Overall, AI aptly reflects common demographic differences in preferences across investment options. These results are robust to changes in the AI model, variations in prompts, and differences in the level of randomness imposed to the AI-generated responses.

To examine the real economic impact of these results, we supplement our analysis of investment preferences with an analysis of actual asset allocations. Specifically, we leverage the asset allocation data from the 2022 Survey of Consumer Finances to estimate demographic differences in allocations to stocks, bonds, and cash, and the extent to which they are captured by AI. We find strong agreement between the demographic patterns in the SCF data and AI-generated allocations: AI correctly reflects the differential allocations across gender (women leave a higher proportion of their assets in cash holdings), age (older individuals allocate more of their total holdings to bonds), and income (high-income individuals allocate more of their holdings to stocks and leave a lower proportion of their assets in cash).

Next, we examine the factors driving investment ratings by drawing insights from the free-form *text* explanations accompanying each rating. First, we observe that risk and return are the two most prevalent themes in both human and AI-generated responses. To quantify a respondent's perception of each investment option along risk and return dimensions, we construct two numerical axes representing an asset's perceived risk and return using a textual embedding technique called Semantic Axis (An, Kwak and Ahn, 2018). The orderings of stocks, bonds, and cash from the AI-generated responses and human survey responses are consistent along both dimensions: stocks are deemed to have the highest return, followed by bonds, then cash; and stocks are deemed to have the highest risk, followed by cash, then bonds.

Critically, the mappings between the risk and return representations of the free-form explanations and the associated numerical ratings are very similar between human participants 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, with the same relationship for AI responses. A one-standard-deviation increase in the human participants' risk perception translates into a 0.56-standard-deviation lower rating. Overall, generative AI accurately reflects the relationship between risk and return encodings and categorical ratings of investment options, which goes well beyond simply documenting existing empirical facts.

We further develop the analysis of free-form explanations by identifying auxiliary

themes impacting stock market participation. Low stock market participation is an enduring concern, especially among female, older, and lower-income individuals (Guiso, Sapienza and Zingales, 2008; Hong, Kubik and Stein, 2004; Van Rooij, Lusardi and Alessie, 2011). We use generative AI's summarization capabilities to extract the main themes (other than risk and return) differentiating explanations for high versus low ratings of stocks, separately for human survey responses and AI-generated responses. The two main themes emerging from the summarization exercise are (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 resulting emotional responses) (Malmendier and Nagel, 2011, 2016).

Generative AI is able to match the demographic differences in these auxiliary factors: men and higher-income individuals are more likely to express familiarity with the stock market, and younger individuals, men, and higher-income earners discuss more positive stock-market experiences. Moreover, when we examine the *correlations* between the four key themes in the free-form explanations of stock ratings—risk, return, knowledge, and experience—we find very similar patterns in human survey responses and AI-generated responses. Both knowledge and past experiences correlate positively with return perceptions and negatively with risk perceptions, and knowledge and experience are positively correlated with each other.

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 (i) replicates the themes that drive these differences, and (ii) identifies how each theme translates into numerical ratings, how the perceptions vary across demographics, and how the themes relate to each other. For example, instead of the generated responses capturing each factor one at a time (e.g., some positive ratings of stocks reflecting high return expectations, and others reflecting high familiarity with the stock market), the large language model reproduces the *correlations* between these factors within-response. This has interesting implications for future research. Our results suggest the potential of generative AI models to capture the sources of heterogeneity across groups (preferences vs. beliefs vs. information, etc.). This is a promising starting point for leveraging generative AI to understand the drivers of heterogeneity more broadly, especially when direct survey evidence may be difficult to obtain. Generative AI's ability to reflect the underlying structure of explanatory themes can make it a powerful tool for hypothesis formation

when it comes to exploring underlying mechanisms.

Finally, we leverage direct-comparison responses—where participants directly rank pairs of investment options—to examine transitivity. AI-generated preference orderings are transitive 98.7% of the time, compared to 84.4% of human preference orderings. This remarkable difference highlights AI's ability to maintain logical consistency in investment preferences, a critical feature for financial decision-making tools. Importantly, we find that most of the intransitivity in human preferences co-occurs with indifference between at least two investment options. Our textual analysis further reveals that a key driver of this intransitivity is a lack of financial knowledge, mediated by participants' tendency to express indifference. This suggests two promising applications for generative AI. First, transitivity of preference orderings is one area in which AI appears to simultaneously preserve heterogeneity and debias human preferences.[2] Second, AI can help uncover reasons for biases from free-form explanations.

Our results contribute to the rapidly growing body of work on the effect of financial technology on the investing landscape. The availability of low-commission 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 investing landscape comprised of increasingly diverse participants. To date, robo-advising has led to greater diversification (D'Acunto, Prabhala and Rossi, 2019), returns (Reher and Sokolinski, 2024), and bias mitigation (D'Acunto, Ghosh and Rossi, 2022) in portfolios. 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 machine learning—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 assuage those concerns, showing that (i) the default model is biased due to the skewed data it was trained on, but (ii) the model can qualitatively capture preference heterogeneities when the model is seeded with demographic information, and

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

most promisingly (iii) the underlying structure of the knowledge base—including factors that drive heterogeneous preferences—resembles true human reasoning. In doing so, we 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 in particular.[3] There is a growing stream of literature showing that AI can be effectively incorporated into financial and other business practices if proper considerations are taken into account (Bybee, 2023; Hansen and Kazinnik, 2023; Kakhbod et al., 2024; Lopez-Lira and Tang, 2023).[4] We build on this work by examining generative AI's ability to capture individual-level heterogeneity in investment preferences and the underlying reasons driving these differences.[5]

Overall, our results underscore AI's promise in financial advising. 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 structures. Future research should 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. If effectively implemented, AI has the potential to enhance financial literacy, reduce behavioral biases, and help individuals make more informed and rational investment choices.

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 and reveals one factor associated with the lack of transitivity in human preferences, and Section 6 concludes.

[3]The introduction of generative AI (proxied by GPT4) and its effects on the labor market and firm performance have been studied by Bertomeu et al. (2023); Brynjolfsson, Li and Raymond (2023); Eisfeldt et al. (2023); Eloundou et al. (2023); and Noy and Zhang (2023).

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

[5]We also offer a methodological contribution by introducing a flexible, low-cost approach to extracting meaningful information from abstract text embeddings in data such as free-form explanations. This approach allows researchers to evaluate text documents along a flexible set of labeled axes beyond commonly considered linguistic attributes such as sentiment, novelty, and complexity (Fedyk and Hodson, 2023; Loughran and McDonald, 2014; Tetlock, 2007).

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 responses come from a survey of a representative sample of 1,272 individuals recruited through the Prolific platform. The survey was conducted in October 2023 and March 2024.[6] 1,264 individuals completed the entirety of the survey, including the demographic questionnaire at the end. The sample of respondents is balanced on age, gender, and income. Specifically, 49.8% of the respondents reported being male, 47.5% identified as female, 2.3% chose "Other" 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.

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 of these questions had three choices (corresponding to the two assets being compared and an option for indifference) and were accompanied by free-form text entry "why?" questions requiring responses

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

with a minimum of 20 characters. The order of the direct-comparison questions 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 toward 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 performed the following cleaning steps on the survey data before commencing the analysis. First, we removed outliers in age and income using an inter-quantile 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. Additionally, we excluded human responses where the participant refused to disclose their gender or identified as non-binary because we lacked 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 dollars.

2.2 AI data collection

We simulate survey data collection using AI to conduct a cleanly identified comparison with human survey data. As in the human survey, we elicit responses to three types of questions:

1. How do simulated AI agents rank each investment option (stocks, bonds, cash)?
2. What is the stated rationale for the ratings (free-form responses)?
3. What is the preference ordering of AI agents among the three options (pairwise comparisons)?

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 convert these multiple choice ratings to numerical ratings from 1 (very negative) to 5 (very positive). For the second type of question, we ask for a short explanation of each rating response using 5 to 10 words. For the third 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." Additionally, we ask each AI agent to report the gender, age, and income of its imagined identity. Online Appendix C presents a sample prompt.[7]

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 "sees" itself and how it behaves when it is *not* instructed to take on any particular identity.

For the sample of AI responses that *do* take demographics into account, we query the AI 1,200 times, seeding it with different demographic characteristics. This sample is meant to mimic the representative sample of participants in our human survey. As such, we specify the median age and income according to the United States census data discussed in Section 2.1, querying AI responses for imagined male or female agents above or below the median in age and income. AI can follow this instruction 100% of the time. Therefore, we have an even split of the data across the 8 demographic groups: males of below-median age and at most median income, males of at least median age and at most median income, etc. The male-to-female ratio of the reported gender (AI agents) is 50-50, 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 dollars. We use the same outlier detection procedure to remove outliers as we applied to the human survey results. 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 dollars.

3 Replicating human rating patterns with AI agents

In this section, we study the promises and limitations of using AI-based approaches to replicate the distribution of investment preferences across demographic groups. We first

[7] 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.

provide descriptive statistics of average ratings of stocks, bonds, and cash across different demographic groups of human survey participants. We then present a naive approach to generate ratings using AI without specifying demographic explanations. In this case, AI agents display a biased representation of the human population–overweighing the preferences of young, wealthy individuals. Finally, we provide a remedy and show that generative AI does correctly reflect human demographic heterogeneities when demographic information is seeded in the prompts.

3.1 Human investment ratings across demographic groups

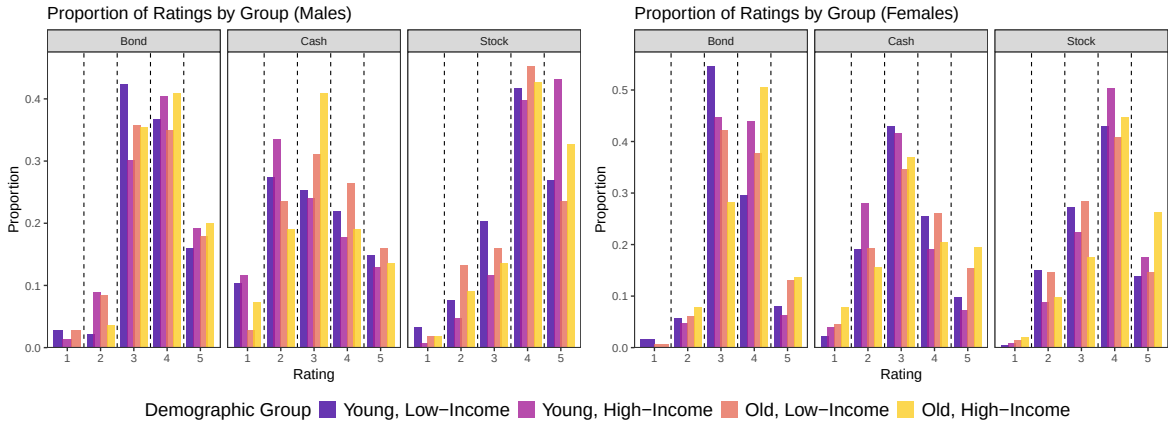


Figure 1: We show the distribution of investment ratings by gender. Panel A (left) shows the distribution of ratings by female survey participants, and panel B (right) shows the distribution of ratings by male survey participants. In addition to the gender split, we consider four subgroups based on age and income, each divided by the median values from the U.S. census. A score of 1 corresponds to a strongly negative rating, whereas a score of 5 corresponds to a strongly positive one.

We consider eight demographic groups, defined by three binary indicators: age, gender, and income. Gender is split into male and female.[8] Age and annual income are sliced at the median based on the 2021 and 2022 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 males) accounts for 17% of the sample, while the smallest groups (old high-income females, old low-income males, and old high-income males) account for 10% each.

[8]As discussed in 2.1, only 2.8% of participants responded with other gender options. To preserve the

Gender	Age	Income	Participants
Female	Old	Low-income	130
Female	Old	High-income	103
Female	Young	Low-income	172
Female	Young	High-income	125
Male	Old	Low-income	106
Male	Old	High-income	110
Male	Young	Low-income	182
Male	Young	High-income	146

Table 1: This table displays the number of human survey participants across demographic groups based on gender, age, and income. 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.

As shown in Figure 1, investment preferences for stocks, bonds, and cash vary across demographic groups. For example, females tend to rate stocks lower than males, and higher-income individuals tend to rate both stocks and bonds more favorably than their lower-income counterparts. Similarly, older individuals assign higher ratings to cash than younger participants.

3.2 Human responses versus default AI agents

We compare the investment ratings of human survey respondents against those generated by default AI agents, based on 150 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. In contrast, default AI agents exhibit similar preferences but assign slightly higher ratings to stocks and bonds, at 4.6 and 4.0, respectively, while rating cash marginally lower at 2.9.

Analyzing the self-reported demographics of AI agents reveals a roughly balanced gender distribution, with 56% identifying as female. However, all agents report annual incomes above the median threshold of $54,000, and only 9 out of 150 (6%) identify as at least 39 years old. This suggests that, by default, AI-generated agents tend to assume the characteristics of young, high-income individuals, potentially introducing algorithmic bias when demographic parameters are not explicitly specified.

We find further evidence of this bias by examining the investment ratings of default

statistical power of our analyses, we exclude this handful of observations.

AI agents. To assess how these ratings align with human survey responses, we divide human participants into eight demographic groups based on gender (male, female), age (below or at least the median), and income (at most or above the median). For each group, we construct a three-dimensional Gaussian distribution based on their ratings of stocks, bonds, and cash, which serve as benchmarks for comparison (Figure 1 presents the one-dimensional rating distribution for each asset class from each group of human survey participants). We then evaluate each simulated AI agent by determining the most likely human demographic group from which its ratings could be 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. Among male groups, 7.333% of AI agents align with the old, low-income category, while 8.667% correspond to the old, high-income category. The largest proportion, 54.667%, is most closely matched with the young, high-income male group. Together, the self-reported demographics and the generated investment ratings provide consistent evidence that default AI agents overrepresent the young, high-income segment of human participants, underscoring the potential for systematic bias when demographic attributes are not explicitly unspecified.

3.3 Adding demographic information to AI prompts

We have shown that generative AI is susceptible to algorithmic bias by disproportionately overrepresenting certain demographic groups within the human sample. To mitigate this issue, we repeat the analysis while explicitly assign demographic attributes to each AI agent when conducting the survey. As detailed in Online Appendix C, we modify the prompt by including the following instruction 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 categorized as 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 the human survey data, we generate 150 responses for each of the eight demographic profiles, yielding a total of 1,200 AI-generated responses.

In the following analyses, we show that introducing demographic information into the prompts results in AI agents producing rating patterns that closely align with those observed in human responses.

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 from each of the eight demographic groups on each of the three investment options. 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. The demographic patterns in the ratings from human survey participants and AI are highly consistent, with a Pearson correlation of 0.73, Spearman correlation of 0.70, and Kendall correlation of 0.57. To compute the statistical significance of these correlation coefficients, we use a bootstrap of 10,000 samples drawn randomly with replacement from the human survey and AI-generated data, respectively. Each bootstrap has the same size as the original 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 responses generated by humans and AI show a Pearson correlation of 0.78, a Spearman correlation of 0.81, and a Kendall correlation of 0.71. All three correlations are 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).

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 coefficients agree. We focus on coefficients that are statistically significant in both human and AI-generated responses and compare the direction of the effects.

Table 3 shows the results: five out of the six significant coefficients are significant in

	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 AI-generated responses 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 standard errors are reported in parentheses.

the *same* direction in both human and AI data. Specifically, older respondents tend to rate cash more highly than younger ones, women tend to rate stocks lower than men, higher-income individuals rate both stocks and bonds higher than lower-income individuals, and higher-income individuals rate cash less favorably than lower-income individuals. The only aspect on which AI responses fail to correctly capture demographic differences in human investment preferences is the relationship between gender and bond ratings: the generative AI model expects women to rate bonds higher than men, whereas male human survey participants place higher ratings on bonds than women. This one-off difference may be attributed to AI reflecting the standard investment allocation tradeoff between stocks and bonds, where men's higher allocation to stocks comes at the expense of bonds (Agnew, Balduzzi and Sunden, 2003).

Overall, Tables 2 and 3 show that AI responses are highly correlated with human surveys, correctly reflecting that women and older investors have lower preferences for stocks than men and younger investors, and that income is a major driver of investment preferences, with higher-income individuals favoring stocks and bonds and being less attracted to cash.[9] This result is especially surprising in light of the evidence in subsection 3.2, which shows that generative AI often assumes the identity of young and high-income (mostly male) individuals. This suggests that the vast majority of information regarding

[9]This alignment between AI agents' and humans' heterogeneous patterns across demographics is robust when we use subsample bootstrapping to compute the confidence intervals of the coefficients. In this case, we find that higher-income individuals rate stocks and bonds higher than their counterparts, and males rate stocks higher than females. AI and human responses are directionally aligned in all four significant coefficients.

	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 correlations that are significant in data generated by both human participants and AI. A plus sign in the second and third columns means that the corresponding demographic is positively correlated with the rated asset. For example, the first row of this table shows that older individuals and simulated AI agents both prefer holding cash more than their younger counterparts.

investing preferences in the model's source material comes from the young and high-income demographic group. And yet, something as simple as specifying the target demographics in the prompt largely undoes this algorithmic bias. The combination of our results highlights the importance of being mindful of potential heterogeneity in financial applications of generative AI, as algorithmic bias is only a concern when the model receives insufficient guidance from the user.

3.4 Asset allocation

In this subsection, we show the external validity and broader applicability of our results by leveraging real household-level asset allocation data from the 2022 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. It collects detailed information on the financial characteristics of U.S. households, including data on income, net worth, balance sheet components, pension plans, and other financial behaviors. We observe the age, gender, and income of each respondent. The SCF dataset is at the household level, which can include multiple income earners (e.g., married couples). In our main specification, we restrict the sample to single-person households; however, the results remain the same if we include all households (with demographics marked by the head of household) and use fixed effects for the number of people in each household. For the asset allocation,

we observe the amount of money in the savings and checking accounts (cash), the amount invested in publicly traded stocks, and the amount invested in bonds. We compute the fraction of allocation in 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—specifying the age, gender, and income of a participant and asking for an allocation across the three asset classes adding up to 100%.

We analyze the demographic differences in allocations, and whether these align between actual human data (the SCF) and responses generated by AI. The significant differences in asset allocation across demographic groups are shown in Table 4, analogous to the demographic differences in preferences reported in Table 3. Older individuals invest more in stocks and bonds, females keep a larger proportion of their assets in cash, and high-income earners invest more in stocks and keep less cash. These patterns are mirrored in the responses generated by AI. Overall, large language models are able to match not only demographic differences in self-reported preferences across asset classes but also how those differences translate to actual asset allocation. This evidence suggests 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: stocks	+	+	✓
old: bonds	+	+	✓
female: cash	+	+	✓
high-income: stocks	+	+	✓
high-income: cash	-	-	✓

Table 4: This table displays the relationship between the portfolio allocation to each asset and demographic characteristics (age, gender, and income). We show the five correlations that are significant in both human response data and the AI-generated data. A plus sign in the second and third columns means that the corresponding demographic is positively associated with the allocation to the corresponding asset.

3.5 Robustness and Additional Analysis

We examine the robustness of our main result to several variations in the large language model. First, we show that the results are the same between the standard GPT4 model and the newer GPT4o model. Second, we demonstrate that the results are robust to variations

in the AI prompt design, specifically considering a practical extension of the prompt: providing the model with access to a small sample of real human survey data. Finally, we show that our results are robust to reducing the randomness in the AI-generated samples by setting a lower temperature hyperparameter.

3.5.1 Robustness to model choice: generating responses with GPT4o

We show the robustness of our main result to the choice of the large language model by replicating the same analysis with a newer state-of-the-art model—GPT4o (released in May 2024).

We use the same prompts as the main analysis and collect the same sample size in each demographic group (150 responses). Tables D1 and D2 in the Online Appendix show that the newer model performs even better than our baseline results: GPT4o responses are more highly correlated with human survey responses (compared to GPT4), and the rating patterns across demographic groups match better. More details on this analysis can be found in Online Appendix D.1. These results demonstrate that our key finding is not restricted to one GPT model, and the alignment between large language models and human surveys is likely to increase as newer, more advanced models are released.

3.5.2 Prompt variation: low-cost few-shot learning

We explore the robustness of our results to varying the input to the large language model by considering an important application. Although running large human surveys can be costly and time-consuming, sometimes researchers have access to a small pilot sample containing some real survey data, which may be too small to draw statistically significant inferences but can inform research design. We show that incorporating this small dataset into the large language model through a low-cost fine-tuning approach called few-shot learning can further improve the alignment between the responses generated by AI and real humans.[10]

We begin with a small random "pilot" sample of 100 respondents from the human survey data (under 10% of the entire sample). We then embed a random set of three of these "pilot" responses from the same demographic group (age, gender, and income) into

[10]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 differ from them in that few-shot learning does not directly change the embeddings produced by the model but rather using prompts to guide it to produce answers that are aligned with our examples.

each prompt when querying AI. Online Appendix D.2 presents the full prompt and a more detailed discussion of the process. We then collect the AI-generated responses like before, with 150 queries for each demographic group. We compare these AI-generated responses to the *remaining* responses from the human survey (i.e., those that were not used as the 100 "pilot" sample). As shown in Tables D3 and D4 in the Online Appendix, the alignment between the few-shot learning results and real human responses is even higher than in our baseline analysis: correlations between demographic group averages reach 80%, and there is directional alignment on all seven significant relationships between demographics and ratings. Therefore, if researchers have access to a small pilot dataset of real human responses, it is beneficial to include this information via few-shot learning when using AI to generate synthetic survey responses.[11]

3.5.3 Variation in temperature

We examine whether our results are robust to variations in the level of randomness in the AI model. We control the randomness through a hyperparameter known as temperature.[12]

Our main analysis employs a relatively high temperature of $T = 0.8$ to introduce diversity in AI-generated responses. To assess the robustness of our findings, we regenerate AI responses with reduced randomness by setting $T = 0.5$. Using the same prompt as detailed in Appendix C, we generate 150 samples for each of the eight demographic groups. We then compute a vector of 24 group-level averages—representing the three asset classes across the eight demographic groups—and estimate their correlation with human survey data. The Pearson correlation between AI-generated and human averages is 0.69, the Spearman correlation is 0.68, and the Kendall correlation is 0.56, all statistically significant at the 1% level. Moreover, as shown in Table D5, these correlations remain significant at least at the 5% level within each asset class.

[11]The pilot sample of 150 responses alone has low statistical power and uncovers only two significant correlations—that older individuals rate cash higher than younger individuals and high-income earners rate stocks higher than low-income earners. Leveraging AI helps overcome the low-power issues of small datasets like this sample, uncovering the full set of significant correlations in the data.

[12]The temperature parameter (T) controls the randomness of text generation by scaling the logits before applying the softmax function. Given a probability distribution over tokens, the modified probability of token i is computed as $P_i = \exp\left(\frac{z_i}{T}\right) / \sum_j \exp\left(\frac{z_j}{T}\right)$, where z_i represents the model's raw logit for token i. Lower values of T (e.g., $T \to 0$) amplify the probability of the most likely tokens, making responses more deterministic and repetitive. Higher values of T (e.g., $T \to 2$) flatten the distribution, increasing diversity but potentially reducing coherence. Typical generative AI applications set T to a number in $(0, 1]$.

To further examine the consistency of AI-generated preferences, we replicate the demographic regressions from our main analysis using the lower-temperature AI data. As reported in Table D6, we find that nearly all coefficient pairs between AI and human data are directionally aligned. Specifically, both older AI agents and human participants exhibit a stronger preference for cash and bonds relative to younger individuals. Similarly, higher-income respondents, whether human or AI, allocate more weight to stocks and bonds while assigning less to cash. The only notable divergence is that male human participants rate bonds higher than their female counterparts, whereas male AI agents rate bonds lower. These patterns are identical to Table 3. Overall, these results reinforce our main conclusion: AI with demographic seeding can effectively replicate human investment preferences across different levels of randomness.[13]

4 Understanding explanations of ratings

In Section 3, we have shown that default large language models suffer from algorithmic bias, overrepresenting younger, high-income individuals in investment preferences. However, when queried for specific demographic groups, AI-generated investment ratings closely align with human survey ratings.

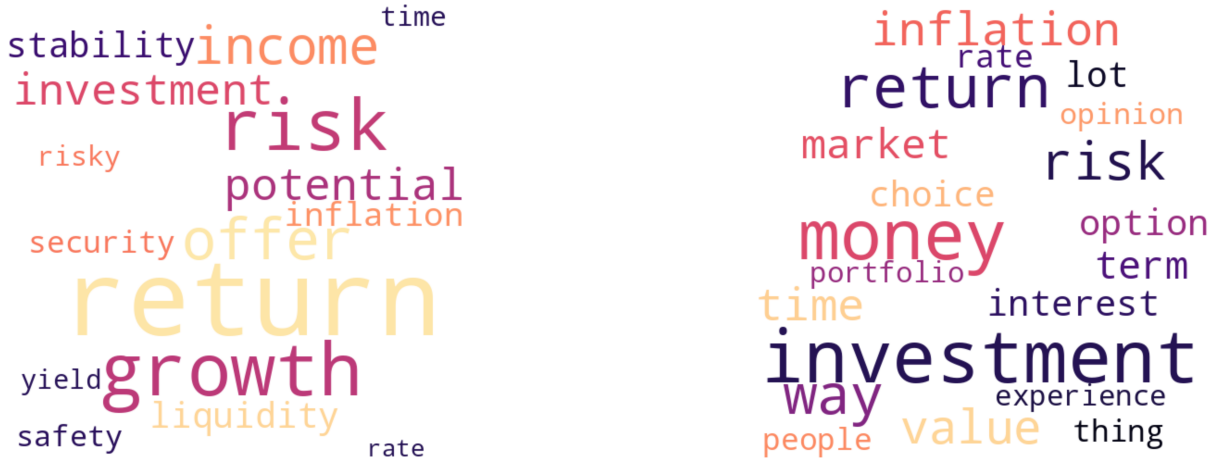
We now extend the analysis to examine whether generative AI captures not only investment ratings but also the reasoning behind them. Specifically, we analyze the free-form justifications accompanying each rating to assess the extent to which large language models replicate the underlying structure of human explanations. By identifying key themes, measuring their association with investment preferences, and exploring their interconnections, we evaluate the degree to which the rationales supporting AI-generated responses mirror human patterns in financial decision-making.

4.1 Most common themes

We begin by conducting a simple word count of all nouns that appear in human and AI responses, separately, to determine the most frequent themes being discussed. Table A1 in the Online Appendix shows the 15 most common nouns in the responses from the human survey (on the left) and the 15 most common nouns in the AI-generated responses (on the right). The human responses use more varied language than those generated by

[13] More details are in Appendix D.3.

AI (the most common term in AI-generated responses—"return"—appears 1,408 times, compared to a third of that for the most common terms in human responses). However, the most common themes in both sets of responses are similar, focusing on investments, risk, and return. Figure 2 offers a graphical illustration of the most common terms using word clouds. There are differences in the terms used—human responses focus on "investment" and "money," whereas AI generates more discussion of "return" and "growth," but the general patterns are similar: both sets of explanations concentrate on financial tradeoffs and rewards.



(a) AI word cloud

(b) Human word cloud

Figure 2: Word clouds of AI and human responses constructed based on the frequency of words in the free-form explanations. We only show words that are used as nouns in the explanation because we are interested in detecting the major themes. Larger words appear more frequently.

4.2 Discussions of risk and returns in human and AI's explanations

Motivated by the main themes emerging from the most frequent terms, we conduct a more rigorous examination of the two main themes in the responses, which correspond to the two principal dimensions of utility functions studied in financial economics: risk and return. First, we construct two semantic dimensions using embeddings.[14] We begin

[14]This 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).

with the embeddings of four principal sentences:
Return: "This asset has very high return." and "This asset has very low return.";
Risk: "This asset has very high risk." and "This asset has very low risk."

We denote these four embeddings as $\mathbf{V}^h_{\text{ret}}$, $\mathbf{V}^l_{\text{ret}}$, $\mathbf{V}^h_{\text{risk}}$, and $\mathbf{V}^l_{\text{risk}}$ (normalized to have unit length), respectively. Embeddings are numerical representations of text, and 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 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 $\mathbf{V}^l_{\text{risk}}$:

$$\mathbf{V}_{\text{risk}} = \frac{\mathbf{V}^h_{\text{risk}} - \mathbf{V}^l_{\text{risk}}}{\text{norm}\left(\mathbf{V}^h_{\text{risk}} - \mathbf{V}^l_{\text{risk}}\right)}.^{15} \tag{2}$$

Next, we extract the embedding of each explanation in the human response data and the AI-generated data. We decompose the meaning of each explanation into two components: return-related and risk-related. Given an embedding vector $\mathbf{emb}_i$ of explanation i, the decomposition can be computed as

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

where $c_{i,r}$ is the projection coefficient of $\mathbf{emb}_i$ onto $\mathbf{V}_{\text{ret}}$ and $c_{i,v}$ is the projection coefficient onto $\mathbf{V}_{\text{risk}}$. Intuitively, $c_{i,r}$ is explanation i's association with high return, and $c_{i,v}$ is i's association with high risk.

We begin the analysis of textual explanations by confirming that explanations generated by AI align with human responses in terms of the relative ordering of the asset classes (stocks, bonds, and cash) along the return and risk dimensions. Both human participants' and AI-generated explanations have, on average, the highest return components when

[15] We find that the two axes ($\mathbf{V}_{\text{ret}}$ and $\mathbf{V}_{\text{risk}}$) are nearly orthogonal with an angle of 70 degrees.

explaining rankings of stocks, followed by bonds, then cash. Similarly, both human and AI-generated explanations project the highest risk when discussing stocks, followed by cash, and then bonds. AI's relative discussion of stocks, bonds, and cash along both risk and return axes is consistent with the verbal explanations provided by human survey participants.[16]

Next, we combine the information from the free-form explanations and the ratings on different investment options generated by humans and AI to estimate the correlation between the perceived risk and return and the stated rating of each investment option across human participants 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 reflected in responses generated by AI. We estimate the following specification:

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

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. $c_{i,,k,r}$ 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) space, likewise standardized to have a mean of zero and a standard deviation of one. $\delta_{i,k}$ is a fixed effect that denotes the type of investment (stocks, bonds, or cash). The coefficient β_1 (β_2) can be interpreted as the importance of return (risk) for humans' ratings of the investment options, β_3 (β_4) can be interpreted as the under- or overstatement of the importance of return (risk) by generative AI, and δ_k are the fixed effect of each investment options k interacted with an indicator of whether this rating is from AI or a human participant.

The results are presented in column (1) of Table 5. Naturally, a higher perceived return-generating potential of an asset in the free-form responses is significantly positively correlated with higher categorical ratings for that asset. The notable finding is that this relationship is exactly the same in the human survey responses and the AI-generated responses. In the human survey, a one-standard-deviation increase in the perceived re-

[16]Some real examples of explanations with high and low projection scores are shown in Table A2 in the Online Appendix.

	Dependent variable:	
	Rating	
return	0.788^{***}	0.789^{***}
	(0.015)	(0.015)
risk	-0.564^{***}	-0.549^{***}
	(0.017)	(0.017)
AI bias: return	-0.010	-0.029
	(0.021)	(0.021)
AI bias: risk	0.003	0.011
	(0.028)	(0.028)
Demographics × AI		✓
Observations	6,348	6,348
R^2	0.556	.560
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 5: Statistical analysis of the association between risk and return projections of free-form explanations and the categorical ratings. This table shows the association based on human data and the differences in the association between human and AI-generated data. The left column has risk-return projections and the interactions between the AI-generated indicator and risk and return as independent variables. The right column shows the coefficients estimated when we include additional controls—interactions between the AI-generated indicator and demographic variables—age, gender, and income.

turn is associated with a 0.79 standard deviation increase in the rating. AI replicates this relationship between ratings and returns, with no significant bias. A one-standard-deviation increase in the perceived return is associated with a 0.78 standard deviation increase in the rating in AI-generated responses, statistically indistinguishable from the corresponding coefficient in the human data. A similarly aligned pattern is observed on the risk side too: each standard deviation increase in risk projection is associated with a 0.564 standard deviation decrease in ratings in the human data and 0.561 in the AI-generated data. In addition, as shown by column (2), these results are relatively unchanged both in direction and magnitude when we also control for demographic variations interacted with the data source in equation 4. Overall, we observe that survey responses generated by AI closely match human responses in terms of the "reasoning" behind the ratings, with very similar relationships between ratings and risk-return factors.

4.3 Other themes in free-form responses

So far, we have examined the two major common themes in free-form explanations of investment preferences, risk and return. We now investigate which other auxiliary themes play a role in the data, and to which extent these agree between human survey data and AI-generated responses. We focus specifically on explanations for stock ratings, to speak to the drivers of low stock market participation, which is a long-standing and significant issue documented by the prior literature (Guiso, Sapienza and Zingales, 2008; Van Rooij, Lusardi and Alessie, 2011).[17]

To do so, we go beyond the most common single terms (unigrams) and use the full text of the responses to capture potential themes conveyed by phrases or sentences. Specifically, we combine *all* of the explanations about a specific investment option (e.g., 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 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 non-positive ratings)*

This gives the main five auxiliary themes in the human response data. Then we use the same prompt to identify the auxiliary themes based on 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 themes from each run are listed in Table A3 in the Online Appendix.

In all summaries, the themes in both the human and AI-generated data prominently feature the topic of "knowledge" (financial literacy and understanding of the stock market). Specifically, human data are summarized by the themes "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

[17] Our main emphasis is on stock market participation given its central role in the literature, but for completeness we also provide analogous analysis of the auxiliary themes for bonds (where knowledge/familiarity emerges as the main auxiliary theme) and cash (where the most consistent theme is the convenience/accessibility of cash) in Online Appendix F.

"Knowledge and Complexity" (second and third run). Beyond knowledge and understanding, one other theme that appears frequently in both human and AI-generated data is "experience" (positive or negative emotional reactions to past experiences with the stock market). Human survey responses include themes such as "Emotional Response" (first run), "Influence of Past Experience" (second run), and "Emotional and Psychological Experience" (third run), while AI-generated responses include themes such as "Emotional and Psychological Responses" (first run) and "Emotional Response and Comfort Level" (second run).

First, we focus on the knowledge dimension. We use the Semantic Axis approach discussed in Section 4.1 to compute each response's relevance score to the knowledge theme. In particular, we define the knowledge dimension using 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).

Following the procedure discussed in Section 4.1, we use the difference between the embeddings of these two sentences as a numerical representation of the knowledge dimension. Then, we project the embedding of each explanation accompanying a given (human or AI-generated) rating about investing in stocks onto this axis to compute its relevance to knowledge about the stock market. Figure 3(a) shows a graphical representation of this projection, in blue for the explanations provided by human participants and in yellow for the explanations generated by AI.

Both projections center just below zero, with tails in both positive and negative directions. AI-generated responses show a clear bimodal distribution and smaller tales, while the human responses are less bimodal and have longer tails. We use an expectation-maximization algorithm with two clusters to group data points to their corresponding clusters, where one cluster (positive) reflects the presence of knowledge regarding stocks, and the second cluster (negative) 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. Mathematically, the clustering algorithm can be written as the following optimization problem:

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.

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 PDF 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.[18]

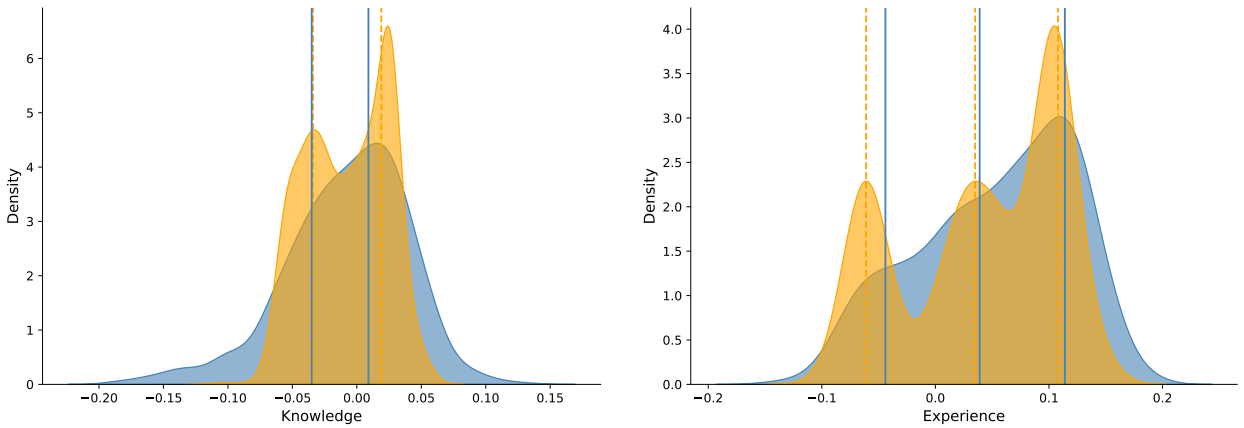
The resulting clusters are displayed in Figure 3(a), with solid blue vertical lines representing human survey explanations and dashed yellow vertical lines for AI-generated explanations. While AI-generated explanations tend to be slightly more positive—cluster centers shift slightly to the right—the 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.

In the first two columns of Table 6, we examine how knowledge levels vary across demographic characteristics, and whether these patterns are correctly reflected by AI. In particular, we consider the cluster label of each agent as its type. This is motivated by the underlying assumption that there are two types of individuals in terms of knowledge about the stock market: the uninformed respondents are randomly drawn from the negative cluster and the knowledgeable respondents are randomly drawn from the positive cluster.

[18]The convergence tolerance we use is 0.0001. In our analyses, all optimizations converge with fewer than 100 iterations. For more details about Gaussian Mixture Models and other finite mixture models, refer to McLachlan, Lee and Rathnayake (2019).

For each data set (embeddings of human responses and embeddings of AI-generated responses), we regress the cluster label (cluster 1 or 2) against the three demographic characteristics: age, gender, and income. The resulting coefficients estimated on the projections of human responses show that the answers of men and higher-income individuals reflect a higher level of knowledge about the stock market. Similar patterns hold in AI-generated data: simulated male responses reflect higher levels of knowledge (with a statistically indistinguishable coefficient to the human data), and simulated responses of higher-income individuals likewise reflect greater knowledge of the stock market (with a more substantial difference than in the human data). Explanations generated by AI are also more likely to project higher knowledge of the stock market from younger individuals, although the association between age and knowledge in the actual survey responses is null.

Second, we conduct similar analyses along the "personal experience" dimension using the difference between the embeddings of the following two sentences: "I have had very good experiences investing in the stock market." (positive experience) and "I have had terrible experiences investing in the stock market." (negative experience).



(a) Distrbution of knowledge projection.

(b) Distribution of past experience.

Figure 3: The density of AI and human responses' relevance to having a high amount of knowledge about the stock market (subfigure a) and positive versus negative past experience (subfigure b). The blue density plot represents the human data, and the yellow plot represents the AI data. The dotted vertical lines represent the centers of the clusters of AI data, and the continuous vertical lines are the centers of the clusters of human data.

Figure 3(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 distribu-

tions, defining three clusters corresponding to more positive experiences, more negative experiences, and neutral experiences. The solid blue vertical lines in Figure 3(b) mark the centers of the three clusters for the experience projections of human responses, and the dotted yellow vertical lines represent the centers of the three clusters for the projections of AI-generated responses.

Each dataset contains three response clusters reflecting different experiences with the stock market. One cluster captures negative past experiences, centered around -0.044 (-0.061) for human (AI-generated) responses. Another reflects mildly positive experiences, centered around 0.039 (0.035), and the third represents very positive experiences, centered around 0.114 (0.108). The distribution of responses across these clusters is similar between human and AI-generated explanations: 23% (24%) fall into the negative cluster, 33% (35%) into the mildly positive cluster, and 44% (41%) into the strongly positive cluster.

We examine how reported personal experiences with the stock market vary across demographics, and whether AI is able to capture those differences. We regress the cluster labels on age, gender, and income of the corresponding human participant or simulated AI agent. The last two columns in Table 6 report the results. Column 1, which reflects actual human responses, shows that younger individuals tend to have more positive experiences with the stock market than older individuals, males tend to have more positive experiences than females, and higher-income individuals tend to have more positive experiences than lower-income individuals. All of the coefficients are significant at the 5% level or more. Column 2 shows that, while the coefficients are different, the directional patterns are the same in AI data: simulated younger individuals, men, and individuals with higher incomes generate explanations containing positive experiences with the stock market.

Together, Table 6 shows that there are strong demographic patterns in the expressions of both knowledge and personal experiences of the stock market and that generative AI is effective at capturing these directional patterns across demographics.

4.4 Text-revealed factor structure

Risk, return, experience, and knowledge are the key themes shaping attitudes toward the stock market. Generative AI replicates human patterns across each of these dimensions. Next, we examine how these themes are correlated and whether generative AI accurately

	Dependent variable:			
	Knowledge		Past 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
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6: This table shows the correlations between age, gender, and income with stock market knowledge (columns 1 and 2) and past experience (columns 3 and 4). The left columns present results for human responses, while the right columns present results for AI responses. Income is scaled in thousands of dollars.

captures their interconnections.

To do so, we analyze the correlations between the four projections computed in the previous sections. For instance, a higher projection on the return dimension reflects a response indicating stronger expectations of stock returns, while a higher projection on the experience dimension corresponds to a response conveying positive past experiences with investing.

Table 7 presents the cross-correlations between these four themes in the *human* investment survey, controlling for demographic characteristics (age, gender, and income). The correlations are statistically significant: both knowledge and positive past experiences are associated with higher return expectations and lower risk perceptions. Moreover, knowledge about the stock market is strongly correlated (0.74) with positive past experiences.

Table 8 replicates this analysis for the *AI-generated* responses, yielding similarly strong and statistically significant correlations. For example, knowledge about the stock market exhibits a correlation of 0.90 with positive past investment experiences. These results suggest that generative AI captures not only individual themes but also the relationships between them, closely mirroring the cross-relationships observed in human survey responses.

Generative AI 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 ensure the correct co-occurrence of themes within responses. Inmportantly, our results indicate that generative AI can replicate not only the ratings and their underlying justifications but also the structural relationships between these justifications, aligning with the patterns found in human survey data.

	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
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 7: This table shows the correlations between the four main themes in textual explanations of human survey participants' stock ratings: risk, return, experience, and knowledge (financial literacy). The correlations are computed with controls for participant gender, age, and income.

5 Are humans and generative AI transitive?

In economic theory, transitivity is a necessary condition for preferences to be deemed rational. In this section, we investigate whether this condition holds for both human respondents and AI-generated preference orderings. Specifically, the third part of our survey experiment asked participants and AI agents to compare three pairs of investment options—stocks versus bonds, bonds versus cash, and cash versus stocks. We leverage these comparisons to illustrate how generative AI can help de-bias human survey responses, providing a proof-of-concept for broader applications.

	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
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 8: This table shows the correlations between the four main themes in textual explanations of AI-generated stock ratings: risk, return, experience, and knowledge (financial literacy). The correlations are computed with controls for participant gender, age, and income.

5.1 Indifference and intransitivity

To test whether each individual's preference ordering satisfies 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-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 $a = b$ or $b = c$ 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 $a, b \in \{cash, bonds\}$.

In Table 9 we document 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. To further investigate the cause of this discrepancy between humans and simulated AI agents, we identify the source(s) of intransitivity in human survey data.

We start by dividing the human sample into responses from men and women. As shown in Table 9, 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.[19]

	AI		Human	
Observations	1042		1074	
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 9: Preference transitivity of human participants and simulated AI agents. In this table, the "diff" columns report the results based on the subset of human participants who did not respond with "indifferent" to any of the three preference questions. The "indiff" columns report the results conditioned on the subset of human participants who responded with "indifferent" to at least one of the three preference questions.

Combining these two observations, we investigate whether the lower proportion of female participants with transitive preference orderings is due to higher incidence of indifference among women. Indeed, we find that 40.3% of male survey participants have

[19]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.

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

5.2 Financial knowledge and indifference toward investment options

In subsection 5.1, we established that indifference between investment options is the clearest indicator of human survey participants violating transitivity. Building on that result, we now analyze the explanations from the human survey data for each pairwise comparison—stocks versus bonds, bonds versus cash, and cash versus stocks—to uncover the text-based reasons underlying expressed indifference.

Similar to section 4, we ask AI to summarize the theme(s) that differentiate two subsets of the data: human participants who responded with at least one indifference and those who provided only strict preferences. We use the 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 Table A4, the main theme that consistently appears in the top 5 themes is knowl-

edge; more specifically, knowledge about bonds and stocks. We use a similar projection approach to the one discussed in Section 4 to measure the level of knowledge reflected by each participant's response. 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).

Then, we combine the comparison responses from each participant into one string and extract its embedding $\mathbf{emb}_i$. We also compute the embeddings of the two prompts ($\mathbf{V}^h_{\text{knowledge}}$ and $\mathbf{V}^l_{\text{knowledge}}$) and estimate the knowledge level of participant i 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:

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

where indiff_i is a binary indicator that denotes whether participant i has expressed at least one indifference between two investment options.

As shown in Table 10, financial knowledge in 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 variable 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 less likely to express indifference between two or more options. Moreover, comparing the t-statistics and McFadden 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 interacting knowledge with demographics adds little incremental explanatory power.

Next, we examine the association between expressed knowledge and transitivity in preference orderings, with a particular focus on whether this association is mediated by indifference toward at least two investment options. We adopt the mediation analysis approach detailed in Tingley et al. (2014), investigating whether participants' financial knowledge is associated with transitive investment choices and whether this relationship

	Dependent variable:		
	indiff		
Knowledge	-0.773^{***}	-0.720^{***}	-0.790^{***}
	(0.075)	(0.076)	(0.277)
Demographics		✓	✓
Demographics × knowledge			✓
Observations	1,074	1,074	1,074
McFadden R^2	0.0871	0.1139	0.1158
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 10: This table shows the association between the level of knowledge reflected by each participant's responses to comparisons between investment options and whether or not the participant is indifferent between at least one pair of investment options. The knowledge level is standardized. 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.

is also linked to the level of expressed indifference toward investment options. To ensure proper numerical representation, the knowledge variable is transformed into a binary indicator, where participants with knowledge above the median are coded as 1 and those below the median are coded as 0.

The first step involves estimating the model for the mediator, where indifference is regressed on knowledge and demographic covariates (gender, age, and income) using a logistic regression:

$$P(\text{indiff}_i = 1) = \frac{1}{1 + e^{-Z_i^{\text{indiff}}}},$$

where $Z_i^{\text{indiff}} = \alpha_0^{\text{indiff}} + \beta_1^{\text{indiff}} \text{knowledge}_i + \beta_2^{\text{indiff}} X_i$. β_2^{indiff} is a vector of coefficients corresponding to the demographic variables X_i which includes age, gender, and income.

The second step involves fitting the model for the dependent variable of interest, where an indicator variable for "having a transitive preference ordering" is regressed on both knowledge and indifference, along with the same demographic controls. This logistic regression examines whether knowledge is associated with transitive investment

choices and whether this relationship is also linked to indifference:

$$P(\text{transitive}_i = 1) = \frac{1}{1 + e^{-Z_i^{\text{transitive}}}},$$

where $Z_i^{\text{transitive}} = \alpha_0^{\text{transitive}} + \beta_1^{\text{transitive}}\text{knowledge}_i + \beta_2^{\text{transitive}}\text{indiff}_i + \beta_3^{\text{transitive}} X_i$. $\beta_3^{\text{transitive}}$ is a vector of coefficients corresponding to the demographic variables X_i, which includes age, gender, and income.

Once the models are estimated, the mediated and direct associations can be computed. The mediated association quantifies the expected difference in the dependent variable of interest (transitive preference ordering) when the mediator (indifference toward investment options) varies according to the independent variable of interest (participants' financial knowledge about investments) while holding the independent variable constant. This is expressed as:

$$\delta(t) = E\left[Y(t, M(t)) - Y(t, M(t_0))\right],$$

where t and t_0 represent specific values of the independent variable knowledge, $M(t)$ is the predicted value of the mediator under knowledge level t, and $Y(t, m)$ is the predicted value of the dependent variable given knowledge level t and mediator value m.

The direct association represents the expected difference in the dependent variable when the independent variable varies, while the mediator is held constant at the value it would take under a specific knowledge level. This is expressed as:

$$\zeta(t) = E\left[Y(t, M(t)) - Y(t_0, M(t))\right].$$

The total association between financial knowledge and transitive investment choices is the sum of the mediated and direct associations:

$$\tau = \delta(t) + \zeta(t),$$

where τ represents the overall association between knowledge and transitivity. These associations are estimated by combining predictions from the mediator and dependent variable models. The mediated association reflects the extent to which indifference toward investment options is associated with the relationship between financial knowledge and transitive investment behavior. In contrast, the direct association captures the part of

this relationship that is independent of indifference.

We use a bootstrap approach with 1000 simulations to calculate confidence intervals for the mediated and direct associations. The mediated association between participants' knowledge level and having transitive preferences is estimated at 0.059, 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 statistically significant at conventional levels.

In summary, our analysis demonstrates that participants' knowledge about investing in stocks and bonds is significantly positively associated with both avoiding indifferences between investment options and exhibiting transitive investment preferences. Importantly, the relationship between knowledge and transitive preferences is primarily mediated by participants not being indifferent about any investment options.

6 Conclusion

We examine how well generative AI can replicate human investment preferences, especially across demographics. Algorithmic bias is becoming an increasingly salient issue with the rise of machine learning in finance. For example, Bartlett et al. (2022) showcase concerns in the credit space, and multiple banks such as Wells Fargo have been sued for using AI models that led to discriminatory lending. Our results examine algorithmic bias in the context of capturing heterogeneous investment preferences. The baseline model does overrepresent young and wealthy individuals, due to imbalanced training data (e.g., discussions of investment on the Internet). 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 factors driving those heterogeneous preferences, as captured by free-form explanations. The model correctly replicates the main themes, how those themes relate to the categorical ratings, how those themes vary across demographics, and even how these themes interact with each other. Furthermore, we find that human participants are more likely than AI to exhibit intransitive preference orderings. This deviation from rationality is associated with a lack of financial knowledge, which results in expressed 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 and its inevitable applications in the investing space 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 takes the first step, showing that generative AI does well in "understanding" human investment preferences, including demographic differences and expressed reasoning. Future research can build on this finding by examining how well generative AI does in guiding heterogeneous human investors toward more optimal allocations. As the investment landscape becomes not only more automated but also more diverse, with growing shares of small and retail investors, the first step—the ability to accurately reflect cross-demographic heterogeneity in investment preferences, attitudes, and experiences—will be a crucial feature for successful applications of AI in finance.

Second, our novel results on the ability of generative AI to replicate the structure of themes in free-form *expanations* showcases the promise of generative AI to uncover underlying mechanisms driving survey responses and elicited preferences. Stantcheva (2020), Ferrario and Stantcheva (2022), Haaland et al. (2024) highlight the importance of open-ended questions to understand people's expectations and reactions to economic policies. Our results suggest a methodological innovation: not only is natural language processing a powerful tool for structuring and analyzing free-form responses, but recent technological advances in generative AI enable researchers to replicate the full survey design, from categorical (close-ended) questions to open-ended explanations, preserving the factor structure of the underlying themes tying responses together. This is especially important in a domain such as investment preferences, where there is substantial preference heterogeneity with very different implications from different *reasons* for such heterogeneity. If differences in preferences are driven by different risk aversion parameters, investment advisers should take this into account; if 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 these underlying reasons and mechanisms, potentially reducing the need to run costly and time-consuming human surveys.

References

Abis, Simona, and Laura Veldkamp. 2024. "The changing economics of knowledge production." *The Review of Financial Studies*, 37(1): 89–118.

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*.

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." Vol. 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. "Measuring what is top of mind." National Bureau of Economic Research.

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

Hochberg, Yael, Ali Kakhbod, Peiyao Li, and Kunal Sachdeva. 2023. "Inventor Gender and Patent Undercitation: Evidence from Causal 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.

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.

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."

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 collect data from AI.
- Section D shows additional robustness analyses.
- Section E contains supplemental materials including a proof of Proposition 1 and an introduction of GPT4.
- Section F shows analyses of themes discussed in explanations of cash and bond ratings.

A Additional Tables

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 A1: Top 15 most frequent nouns in AI and human responses after removing stop words. The ones colored green are characteristics used to describe an investment option.

	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 A2: This table shows some 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 among the five lowest-rated markings of the corresponding investment option.

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

Table A3: Additional themes from AI and Human Responses. We collected 3 batches of 5 themes each by using AI agents' and human participants' explanations of their ratings of investing in stocks. More specifically, we asked for 5 themes (not risk and return) that differentiate low and high ratings.

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

Table A4: Additional themes from Human Responses. We collected 3 batches of 5 themes each by using human participants' explanations of their direct pairwise preferences between stocks, bonds, and cash. More specifically, we asked for 5 themes (not risk and return) that differentiate people who expressed indifference between at least 2 options versus ones that did not.

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.]

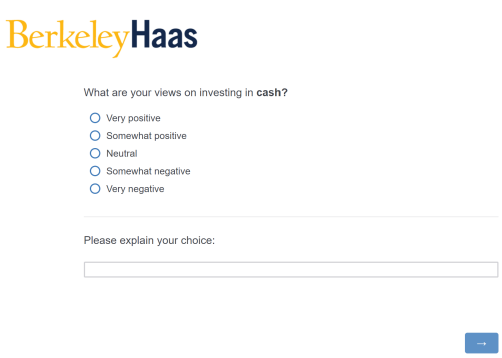
[Questions with comparisons of stocks versus bonds, stocks versus cash, and bonds versus cash—following the example in Figure 1(b)—with the order of the questions and the order in which the options are listed within each question both randomized.]

[Demographic questions screen:]

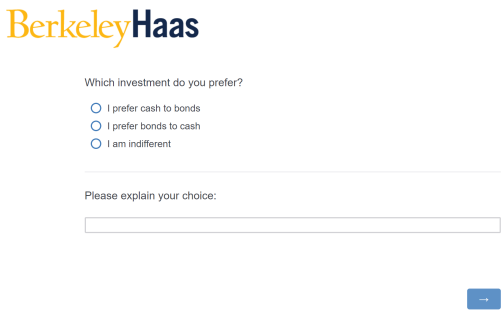
What is your gender? [Options: Male; Female; Non-binary / third gender; Prefer not to say]

What is your age?

What is your gross annual income?



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



(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

We detail the prompts and sample sizes we use to collect data from GPT4. To collect GPT4o-agents' asset allocations, we use the following type of 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 vary the gender, age, and income of each agent: male versus female, above or at most 39 years old, and above or at most 54 thousand dollars of annual income.

We collect 300 samples in each of the 8 demographic categories to form a merged dataset containing 2400 responses. Similar to the main analysis, we drop data points based on the 1.5 IQR rule where data points are considered outliers if their age or income is more than 1.5 inter-quartile range away from the median. We conduct the same outlier removal step for the human survey data (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 Correlation between GPT4o-generated responses and human survey responses

	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)

Table D1: This table reports the Pearson, Spearman, and Kendall correlations between GPT4o-generated responses 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 run the same tests as in subsection 3.3 to assess the correlation between ratings generated by GPT4o and human survey responses. Comparing the results in Table D1 against Table 2, we see that all correlations between GPT4o-generated responses and human survey responses are at least as high as the ones between GPT4 and humans. This shows that the result—that large language models can produce financial survey results that reflect the demographic patterns of human survey respondents—is generalizable beyond a specific GPT model and that the alignment only improves with newer large language models.

Next, following the same steps as in Section 3.3, we use a regression model to examine the directional differences in investment preferences of demographic groups separately for human and AI-generated responses. Significant correlations in both GPT4o-generated and human survey responses are shown in Table D2. Six out of seven significant correlations directionally agree. In both human survey responses and GPT4o-generated responses, 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.

	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 relationship between ratings of the investment options (stocks, bonds, and cash) and demographic characteristics (age, gender, and income). We show the 7 correlations that are significant in both human response data and the GPT4o-generated data. A plus sign in the second and third columns means that the corresponding demographic is positively correlated with the rated asset. For example, the first row of this table shows that older individuals and simulated GPT4o agents both prefer keeping cash more than their younger counterparts.

Overall, we show that similar to GPT4, GPT4o produces rating responses to financial survey questions that are highly correlated with real human responses, reflecting similar patterns across different demographic groups.

D.2 Few-shot learning

We reanalyze our main results using another prompt variation, exploring a practical extension of our approach. Acquiring a large, up-to-date human survey dataset is costly and sometimes time-consuming; however, it is much easier to run a small pilot study to acquire a relatively small number of human samples. Here, we study whether we can use few-shot learning to incorporate a small human survey data set when asking the AI to generate more synthetic data.

Few-shot learning is a computer science technique used to query generative language models. In particular, we embed a few examples (few-shot) in the prompt when submitting the query (Brown et al., 2020). This is helpful when we want to inject domain

knowledge into the AI on a subject that may be sparse in its training data, and it can be viewed as an alternative to fine-tuning – with a lower computational cost.

In our case, the few-shot prompts are of the following type:

> *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: 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 use the following steps to conduct the analysis:

1. Randomly select 100 samples from the cleaned (outlier-removed) human survey responses.
2. Save the rest of the human data as testing data.
3. Subset the 100 randomly selected samples into 8 demographic groups defined in subsection 3.3.
4. Generate 150 responses using AI for each demographic group:
    - Randomly choose 3 examples from the corresponding demographic subset without replacement.
    - Embed the examples in the prompt.
    - Query AI and save the response.
5. Compare the responses generated by AI using few-shot learning with the test human data set.

First, we reproduce the correlation table (table 2) with the few-shot learning data. As shown in Table D3, the correlation coefficients are systematically higher than the ones without few-shot learning, except for stocks. In particular, when we pool together the ratings for stocks, bonds, and cash, the correlation coefficients are significantly higher when we use few-shot learning. Furthermore, the correlations' significance levels are improved for ratings on bonds, which are the lowest in Table 2. All correlations using few-shot learning are positively significant at the 1% level.

Next, we check whether the results are more similar in terms of rating heterogeneity across demographic groups. More specifically, we replicate Table 3 with few-shot learning data. As shown in Table D4, all 7 coefficients significant in both datasets are directionally aligned: older individuals rate bonds and cash higher than younger ones, females rate stocks lower but cash higher than males, and high-income earners rate stocks and bonds higher while rating cash lower than their counterparts.

Overall, we observe that when we have a small number of real human survey responses, using few-shot learning can improve the correlation between the AI-generated ratings and real human ratings (out-of-sample) and better align the heterogeneity of ratings across demographic groups.

	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 AI-generated responses using few-shot learning 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.

	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 relationship between ratings of the investment options (stocks, bonds, and cash) and demographic characteristics (age, gender, and income). We show the 7 correlations that are significant in both human response data and the AI-generated data with few-shot learning. A plus sign in the second and third columns means that the corresponding demographic is positively correlated 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 responses with lower randomness

We run the same tests as in subsection 3.3 to assess the correlation between ratings generated by AI and human survey responses when the temperature hyperparameter is set to 0.5. Comparing the results in Table D5 against Table 2, we see that all correlations between AI-generated responses and human survey responses are significant at the 5% level. This shows that the result—that AI with demographic seeding can produce financial survey

	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.

results that reflect the demographic patterns of human survey respondents—is replicable with different temperature hyperparameters.

Next, following the same steps as in Section 3.3, we use a regression model to examine the directional differences in investment preferences of demographic groups separately for human and AI-generated responses at a lower randomness level. Significant correlations in both GPT4o-generated and human survey responses are shown in Table D6. Five out of six significant correlations directionally agree. In both human survey responses and GPT4o-generated responses, 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 main analysis with a temperature of 0.8, the only source of disagreement is the differential rating of bonds across genders.

Overall, we show that similar to GPT4, GPT4o produces rating responses to financial survey questions that are highly correlated with real human responses, reflecting similar patterns across different demographic groups.

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:*

	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 relationship between ratings of the investment options (stocks, bonds, and cash) and demographic characteristics (age, gender, and income). We show the 6 correlations 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 correlated 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.

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

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 lastly, we describe the improvement of GPT4 (and 3.5) compared 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. However, GPT uses a variant of the Transformer architecture with only the decoder component. This means it 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.

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.

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

where $\mathrm{PE}_{(pos,2i)}$ are the positional embeddings of tokens at even positions, $\mathrm{PE}_{(pos,2i+1)}$ are the positional embeddings of tokens at 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. By incorporating positional encoding, GPT ensures that the model can differentiate between tokens based on their position, allowing it to capture sequential dependencies effectively.

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 other tokens in the sequence, enabling it to understand the context and generate text accordingly.

Following the self-attention layers, each position in the sequence passes through a position-wise feedforward 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. The position-wise feedforward networks help refine the representation of each token in the sequence, incorporating both local and global context information.

Furthermore, to stabilize training and facilitate the flow of gradients, GPT incorpo-

rates layer normalization and residual connections after each self-attention layer and position-wise feedforward network. Layer normalization normalizes the activations 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 all of the tokens in the text. More specifically, the raw input to the attention mechanism is a piece of text T. Then the text is broken into sub-word tokens in a parsing process called tokenization. This set of tokens is predefined such 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 whole text. The result is a set of naive embeddings

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

where t_i is the embeddings for token i, "[BOS]" (beginning of sentence) is a special token that is used to denote the start of a sentence, and "[EOS]" (ending of sentence) is a special token denoting 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.

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

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

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

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

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

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 of 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 where 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 main-

tains 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 following list shows some of the sources used to conduct pretraining for the base model for GPT4:

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 a multitude of 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 explanations of cash and bonds ratings

In this section, we use AI to extract themes beyond risk and return in the explanations of ratings of cash and bonds, analogous to the analysis conducted in Section 4.3 for stocks. We leverage 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 discovered 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 which represents 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 1(a), the distributions of the embeddings of human and AI-generated responses both have three humps. Therefore, we use 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 mildly 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).

Table F1 (first two columns) reveals demographic variations in bond market knowledge. 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: while

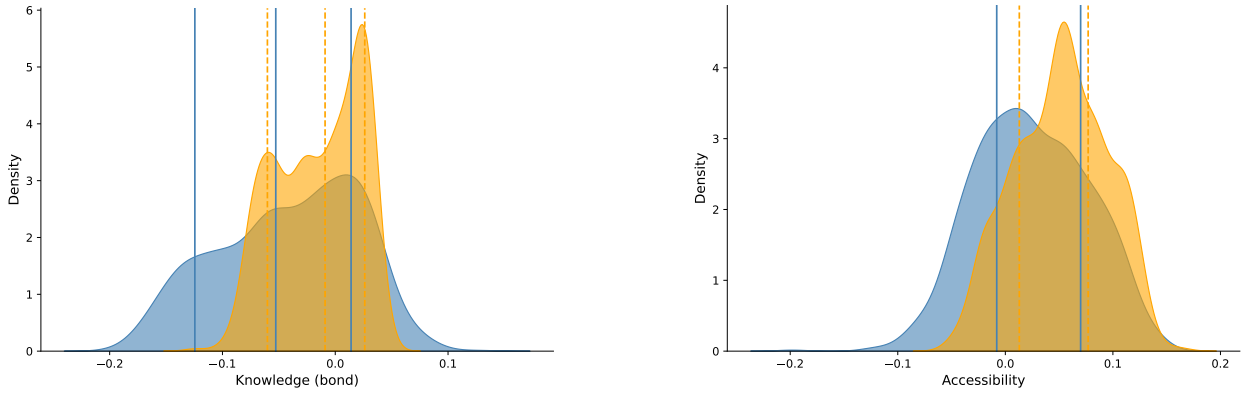
human men express significantly more knowledge about the bond market than women, simulated AI agents tend to express greater knowledge about the bonds market when they are female, although this difference is only marginally significant (at the 10% level).

F.2 Keeping cash

Next, we analyze the themes of the explanations corresponding to positive versus negative attitudes toward holding cash. We find that the most consistent theme in these responses other than risk and return is "accessibility," or the convenience of holding cash, including its liquidity. Some individuals rate cash highly because it is readily accessible at any time. 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."

As shown in Figure 1(b), the distributions of the embeddings of human and AI explanations both have two humps. Therefore, we use a mixture of two Gaussian distributions to cluster the responses. The human distribution contains one positive cluster (centered at 0.070) and one nearly neutral cluster (centered at -0.008 and containing 59% of the participants). The AI distribution also contains one positive cluster (centered at 0.077) and one nearly neutral cluster (centered at 0.013 and containing 39% of the simulated AI agents).

The third and fourth columns in F1 show the demographic variations associated with differences in agents' level of consideration of the high accessibility of cash. Older and lower-income AI agents care more about the high accessibility of cash. However, human participants do not show strong demographic differences in terms of the importance assigned to the accessibility of cash.



(a) Distribution of knowledge.

(b) Distribution of accessibility.

Figure F1: The density of the relevance of human and AI-generated explanations of bond ratings to knowledge about investing in bonds (subfigure a) and cash ratings to caring about the high accessibility of cash (subfigure b). The blue density plot represents the human data, and the yellow plot represents the AI-generated data. The continuous vertical lines mark the centers of the two clusters of human responses, and the dotted vertical lines represent the centers of the two clusters of AI-generated data.

	Knowledge (bonds)		Accessibility of 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

Note: *p<0.1; **p<0.05; ***p<0.01

Table F1: Correlations between age, gender, and income and participants' outcomes. The first two columns report human participants' and AI agents' knowledge about the bond market, while the last two columns report their level of consideration of cash accessibility. Income is scaled in thousands of dollars.