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

Applications of artificial intelligence (AI) in finance have been met with concerns about algorithmic bias, following issues observed in domains such as medical treatment and lending. We ask whether AI models accurately capture investment preferences across demographics. We elicit investment preferences from over 1,200 survey participants and compare the data directly to investment ratings generated by OpenAI's ChatGPT (GPT4). We find that ChatGPT predicts investment preferences with high accuracy across demographics. Specifically, ChatGPT correctly predicts that women rate stocks lower than men, older individuals prefer holding cash, and higher incomes are associated with higher ratings for stocks and bonds. Moreover, free-form responses from ChatGPT focus on the same aspects as human free-form responses. Most common themes in both responses are “risk" and “return," and "knowledge" and "experience" play an important role for stock market participation. One difference is that ChatGPT responses are almost always transitive, whereas human responses are more prone to violating transitivity, especially when expressing indifference. Overall, the use of AI in finance offers a promising direction for augmenting human surveys in preference elicitation, with important applications for areas such as robo-advsing.

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

JEL Classification: C1, G10, G11, G12.

[†] This 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).

1 Introduction

The financial services industry is becoming increasingly automated, with robo-advising (automated investment advice) increasing more than tenfold in the past decade. So far, this has broadly been considered a positive trend. Whereas traditional wealth management was restricted to the richest customers, robo-advising has increased access and benefitted a wider range of investors by improving their diversification, reducing portfolio risk, and mitigating behavioral biases such as trend chasing and home bias (D'Acunto, Prabhala and Rossi, 2019; Reher and Sokolinski, 2024; Rossi and Utkus, 2021). Recent advances in artificial intelligence (AI), including large language models, have the potential to further transform and automate sectors such as the financial services industry (Abis and Veldkamp, 2024; Babina et al., 2024).

However, the effects of these new advances are directionally ambiguous and potentially heterogeneous across investors. On the one hand, advances in AI can lead to efficiency gains and improved performance, as observed in audit (Fedyk et al., 2022) and sell-side analysts (Cao et al., 2021). On the other hand, AI models have suffered from "algorithmic bias"—the tendency to perform better for some demographic groups than others—in a number of areas including medicine (Kadambi, 2021; Zou and Schiebinger, 2018) and image recognition (Buolamwini and Gebru, 2018). Concerns regarding algorithmic bias are especially relevant in the financial domain, where investors display a large gender imbalance (Barber and Odean, 2001). 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). If AI models are trained on imbalanced groups—for example, young men posting on platforms such as StockTwits and Seeking Alpha—then these models may fail to correctly reflect investment preferences of other demographic groups.

In this paper, we directly assess the extent to which state-of-the-art generative AI models (as proxied by OpenAI's GPT4) are able to match differences in investment preferences across three key demographic characteristics shown to predict stock market participation in the prior literature (e.g. Agnew, Balduzzi and Sunden, 2003; Hong, Kubik and Stein, 2004): income, gender, and age. To do so, we first run a real-world custom survey eliciting investment preferences of a representative sample of 1,272 participants. The survey consists of three components: (i) we ask for 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) we ask for a free-form written explanation of each rating; and (iii) we ask participants to compare each pair of investments (stocks vs. bonds, stocks vs. cash, and bonds vs. cash), choosing either the first option, the second option, or "indifferent."

We then query GPT4 with the same survey questions and demographics, with 1,200 simulated runs. The respondents of the human survey (and the simulated GPT4 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 compare the human survey responses and the GPT4-generated responses along three dimensions. First, we analyze the overall ratings each human respondent (or simulated GPT4 agent) assigns to each investment option, examining how the human ratings vary across demographics and the extent to which these patterns are captured by GPT4-generated responses. Second, we take a deeper look at the reasonings (free-form explanations) behind the ratings to study whether GPT4 is able to capture the main themes in the justifications that human participants provide for their investment choices. Third, we discuss one difference, and potential advantage of using GPT4 to model investment preferences: GPT4 responses almost always have transitive preference orderings, whereas human responses are prone to violations of transitivity when expressing indifference.

Our analysis begins by examining the similarity between ratings of investment options generated by GPT4 agents and provided by human participants—across 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 at least as old as the median age, we put him in the group of older high-income (wealthy) men. We construct vectors of human survey participants' ratings and GPT4-generated ratings as the 24 average numerical ratings from each demographic group for each investment option. These two vectors show very high correlations: a Pearson correlation of 0.73, a Spearman correlation of 0.70, and a Kendall correlation of 0.57, all significant at the 1% level. We then repeat the analysis separately for each investment option, with vectors of eight average ratings across the eight demographic groups. The analysis shows positive and significant correlations between human survey responses and GPT4-generated responses across the eight demographic groups within each of the three investment options. Specifically, for bonds, the Pearson correlation is significant at the 1% level, while Spearman and Kendall correlations are significant at the 5% and 10% levels, respectively. For stocks and cash, all correlation types range between 0.65 and 0.81 and exhibit significance at the 1% level. Thus, there is a robust positive correlation between the averages of GPT4-generated ratings and human ratings, indicating the reliability of GPT4 in evaluating investment options across demographic groups.

Looking specifically at the differences in the ratings across demographic groups, we confirm that male human participants rank stocks (bonds) higher (lower) than female participants, 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 GPT4-generated responses, which predict higher stock ratings for men, higher stock and bond ratings for higher-income individuals, lower cash ratings for higher-income individuals, and higher cash ratings for older individuals. Overall, GPT4 aptly reflects common demographic differences in preferences across investment options.

Next, we understand the reasons behind these ratings by drawing insights from the free-form *text* explanations accompanying each rating from human survey participants and simulated GPT4 agents. First, we observe that risk and return are the two most prevalent themes in both human responses and GPT4-generated responses. To quantify each 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 natural language processing technique called Semantic Axis (An, Kwak and Ahn, 2018). We use the two axes to interpret each explanation's relevance to risk and return. We observe that the orderings of stocks, bonds, and cash from simulated GPT4 agents and humans are consistent along both dimensions, with stocks deemed to have the highest return, followed by bonds, then cash; and stocks deemed to have the highest risk, followed by cash, then bonds. GPT4-generated responses tend to be more extreme along both dimensions than human survey responses: compared to human responses, GPT4-generated responses encode higher risk and return for stocks, lower risk for bonds, and lower risk and return for cash. However, 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 GPT4-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 GPT4 responses. In the risk space, a one-standard-deviation increase in the human participants' risk perception translates into a 0.56-standard-deviation lower rating, and the effect size is slightly larger (by 0.05) for GPT4-generated responses. Thus, GPT4-generated responses are about 10% more risk-averse than human participants, but otherwise the relationship between risk and return encodings and categorical ratings of investment options are very similar across actual survey responses and simulated GPT4 responses.

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 for 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 that accompany high ratings of stocks from those accompanying low ratings of stocks, separately for human survey responses and GPT4-generated responses. The two main themes emerging from the summarization exercise are (i) knowledge and understanding of the stock market, and (ii) personal experiences (and resulting emotional responses) with investing in the stock market. Prior literature has shown that financial literacy is a major driver of stock market participation (Van Rooij, Lusardi and Alessie, 2011) and that personal experiences shape attitudes to inflation and risk-taking (Malmendier and Nagel, 2011, 2016). Our comprehensive textual analysis of survey responses confirms that these two themes are key drivers mentioned in individual explanations of positive versus negative attitudes towards the stock market.

We structure the “knowledge" and “experience" themes by embedding the responses in each of these dimensions, analogously to the embeddings we considered for risk and return. We then build a Gaussian Mixture Model to cluster the “knowledge" embeddings into two clusters, corresponding to familiarity with the stock market and an absence of knowledge about the stock market. We likewise cluster the “experience" embeddings into three clusters, with one cluster representing negative experiences, one representing neutral to mildly positive experiences, and one representing strongly positive experiences with the stock market. We find that human survey participants' and GPT4-generated explanations feature similar proportions of each cluster. For example, 23% of human participants' explanations reflect negative experiences with the stock market, compared to 24% of GPT4-generated explanations. Furthermore, the demographic differences in both the knowledge and experience dimensions are very similar in human survey responses versus GPT4-generated responses. Men's explanations are more likely to reflect self-reported familiarity with the stock market than women's explanations, and higher income is likewise associated with greater knowledge and familiarity with the stock market, in both actual survey responses and GPT4-generated responses. Younger individuals' answers reveal more positive experiences with the stock market than older individuals' answers, men discuss more positive stock market experiences than women, and higher-income individuals have more positive experiences than lower-income individuals, both in actual survey responses and in GPT4-generated responses. Thus, GPT4 appears to correctly capture the directionality of the "reasoning" behind attitudes toward the stock market across multiple dimensions, from the canonical risk and return considerations to more subjective experiential aspects such as knowledge and personal experience.

Finally, we identify a difference, and a potential advantage of using GPT4 to simulate survey responses: GPT4 agents almost always follow the transitivity axiom of rational preferences, while human survey participants sometimes do not. To examine the transitivity property within preference orderings, we focus on relative-comparison responses, where participants directly rank pairs of investment options by either choosing one of them as better or expressing indifference between the two options. We prove a set of necessary and sufficient conditions for a participant to have a transitive preference ordering given the format of the data. Applying these conditions, we find that GPT4 agents' preference orderings are almost always transitive, but human participants have transitive preferences less often. We further explore the sources of intransitivity within the human dataset. First, we observe that male human participants have transitive preference orderings significantly more often than female participants. Second, when human survey participants have a strict preference over each pair of options, almost all of them exhibit transitive preferences. However, when at least one "indifferent" response is present, only 72.0% of participants maintain transitivity. Combining these two observations, we find that a larger percentage of female participants report at least one "indifferent" response, and among participants with no "indifferent" responses, men are statistically more likely to exhibit transitive preferences than women. We find that most of the discrepancy in transitivity between human survey participants and simulated GPT4 agents can be explained by these two factors—gender and indifference.

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 advances in technology will impact the investing landscape of an increasingly diverse set of participants. To date, robo-advising has had positive effects on diversification (D'Acunto, Prabhala and Rossi, 2019), returns (Reher and Sokolinski, 2024), and mitigation of biases (D'Acunto, Ghosh and Rossi, 2022). However, the directional effects in terms of both portfolio changes and welfare gains have been unevenly distributed, for example, across 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 the concern 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 in the domain of investing, large language models such as OpenAI's GPT4 correctly reflect directional differences in investing preferences across the demographic characteristics (gender, age, and income) both in numerical ratings and in the free-form explanations behind those ratings.

In doing so, we also contribute to the emerging literature on the effects of AI on different sectors of the economy, with our work speaking specifically to the financial services sector. Veldkamp (2023) and Abis and Veldkamp (2024) spotlight the critical role of data and big data technologies in the modern economy, noting a shift away from labor-intensive processes. The introduction of generative AI (proxied by GPT) and its effects on the labor market and firm performance have been studied by Bertomeu et al. (2023); Brynjolfsson, Li and Raymond (2023); Eisfeldt, Schubert and Zhang (2023); Eloundou et al. (2023); Noy and Zhang (2023). The extant evidence on the effects of GPT technology, particularly in finance, presents a mixed picture. Li, Tu and Zhou (2023) and Bybee (2023) point to the limitations and biases of large language models, indicating potential challenges in their application for professional forecasting. Conversely, Lopez-Lira and Tang (2023) offer a more optimistic view, suggesting that despite errors, large language models may offer valuable predictive abilities beyond human forecasting, underscoring their potential instrumental value in professional settings. We contribute to this literature by analyzing and experimenting with how generative AI can be used to design financial surveys, and whether its use generates substantial bias relative to identically executed human surveys. We document that generative AI is able to reflect similar demographic patterns to human surveys in both elicited investment preferences and accompanying explanations, and that generative AI has an additional advantage of providing responses that do not violate transitivity, even when expressing indifference. Overall, these results portend well for the potential novel application of generative AI in predicting individual (and heterogeneous) investment preferences.

The remainder of the paper proceeds as follows. We describe the data collection procedures for both the human investment survey and the GPT4 prompts in Section 2. We present the main analysis of the consistency between investment ratings in the human survey and GPT4-generated responses in Section 3 and then analyze the explanations of the ratings in Section 4. Section 5 examines the transitivity of relative comparisons by human survey participants and GPT4-generated responses, and Section 6 concludes.

2 Data Collection: Human and GPT4 Investment Preferences

We describe the methodology for collecting human investment preferences across demographics and analogous GPT4-generated responses for our benchmarking exercise.

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.[1] 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. Participants' compensation for participating in the survey corresponded to, on average, $15.45/hour.

The survey asked the respondents to rank 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 rank 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. Panel A of Figure C1 in the Appendix displays an example single-rating question.

Then, the respondents faced three relative-comparison questions: whether they prefer stocks or bonds, whether they prefer stocks or cash, and whether they prefer bonds or cash. Each of these questions had three choices (corresponding to the two assets being compared and to indifference) and were accompanied by free-form text entry “why" questions requiring responses with a minimum of 20 characters. The order of the relative-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

[1] The administered survey was identical 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.

asked whether they prefer "bonds or stocks") to avoid biasing the participants towards any options on aggregate. Panel B of Figure C2 in the Appendix shows an example relative-comparison question. The exact instructions of the survey are included in Appendix B.1.

We performed the following cleaning steps on the survey data before commencing the analysis. First, we removed clear outliers based on self-reported information, such as those reporting an annual income above 500 thousand dollars while participating in an online survey with relatively low compensation for such high-income levels. We removed the outliers in age and income using an inter-quantile range (IQR)-based rule. More specifically, a data point is considered an outlier if the participant's age or income is at least 1.5 IQRs higher or lower than the median. Furthermore, throughout the analysis, 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 (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 GPT4 data collection

We simulate survey data collection using GPT4 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 GPT4 agents rank each investment option (stocks, bonds, cash)?

2. What is the stated reasoning for the ratings (free-form responses)?

3. What is the preference ordering of the GPT4 agent among the three options (pairwise comparisons)?

For the first type of question, we give each simulated GPT4 agent five choices, analogous to 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 GPT4 agent to report the gender, age, and income of its imagined identity. Appendix B.2 presents a sample prompt.

We query GPT4 responses 1,200 times, seeding it with different demographic characteristics. We specify the median age and income according to the United States census data discussed in Section 2.1, querying GPT4 responses for imagined agents that are above/below the median in age and income and either female or male. GPT4 can follow this instruction 100% of the time. Therefore, we have an even split of the data across the 8 demographic groups: males with below median age and at most median income, males with at least median age and at most median income, etc. The male-to-female ratio of the reported gender (GPT4 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 1000 dollars. We use the same outlier detection procedure to remove outliers in the data as we applied in the human survey. The IQR of the GPT4-generated data is smaller than the human-generated ones, indicating that GPT4-generated responses contain fewer extreme outliers. After processing, we retain responses from 1,042 simulated GPT4 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 Investment preferences across demographics

The first question we address is “how similar are GPT4-generated ratings of investment options to human ratings?” In particular, we observe the ratings of stocks, bonds, and cash across eight demographic groups: males with below median age and at most median income, males with at least median age and at most median income, males with below median age and above median income, males with at least median age and above median income, females with below median age and at most median income, females with at least median age and at most median income, females with below median age and above median income, and females with at least median age and above median income.

On aggregate, both human and GPT4 responses rate stocks the highest, followed by bonds, followed by cash. The human survey ratings are, on average, 3.8 for stocks, 3.6

for bonds, and 3.1 for cash. The average GPT4 ratings are 3.7 for stocks, 3.7 for bonds, and 3.2 for cash. There is some heterogeneity between the ratings of different demographic groups, consistent with the literature on demographic predictors of stock market participation. For example, in the human survey, women rate stocks, on average, as 3.6, compared to 3.9 from men. Importantly, similar differences are reflected in the GPT4 responses, which accurately capture preference heterogeneity across demographics.

To assess the similarity between human and GPT4 ratings, we start by computing 24 average numerical ratings from each set of responses (human survey and GPT4): 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 GPT4-generated ratings across these 24 groups. Table 1 shows the corresponding Pearson, Spearman, and Kendall correlations in the top row.[2] The demographic patterns in the ratings from human survey participants and GPT4 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 GPT4 datasets, respectively. Each bootstrap has the same size as the original data. The bootstrapped standard errors are reported in parentheses in Table 1. All three correlation coefficients are highly statistically significant at the 1% level.

We also conduct a similar analysis separately within each asset class, focusing on the eight average ratings (across demographic groups) for stocks, bonds, and cash. These correlations are also very high. For stocks, the differential ratings across demographic groups from human and GPT4-generated responses 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 human and GPT4 ratings for cash show a Pearson correlation of 0.77, Spearman correlation of 0.64, and Kendall correlation of 0.64 across the eight demographic groups, all significant at the 1% level. The correlations for bonds are lower but still statistically significantly, with a Pearson correlation of 0.58 (significant at the 1% level), a Spearman correlation of 0.45 (significant at the 5% level), and a Kendall correlation of 0.27 (significant at the 10% level).

Overall, these results show that GPT4 agents' ratings of stocks, bonds, and cash across

[2] For a detailed definition of each of the correlation coefficients, refer to Appendix Section C.4.

	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 1: This table reports the Pearson, Spearman, and Kendall correlations between GPT4-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.

different demographic groups are highly correlated with real heterogeneity of human ratings across demographic groups. This correlation is robust to using continuous values (Perason correlation) and rank orderings (Spearman and Kendall correlations).

To further examine the exact demographic patterns reflected in human and GPT4-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 this regression for human and GPT4-generated data separately to assess the extent to which the coefficients agree. In particular, we focus on coefficients that are statistically significant in both human and GPT4 responses and assess whether the direction of the effect agrees.

Table 2 shows the results: five out of the six significant coefficients are significant in the *same* direction in both human and GPT-4 data. Specifically, older individuals 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 GPT4 responses fail to correctly capture demographic differences between human investment preferences is the relationship between gender and bond ratings: GPT4 expects women to rate bonds higher than men, whereas male human survey participants actually place higher ratings on bonds than women. This one-off difference may be attributed to GPT4 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 1 and 2 show that GPT4 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.

	GPT4 direction	Human direction	Agreement
old: cash	+	+	✓
female: stocks	-	-	✓
female: bonds	+	-	×
high-income: stocks	+	+	✓
high-income: bonds	+	+	✓
high-income: cash	-	-	✓

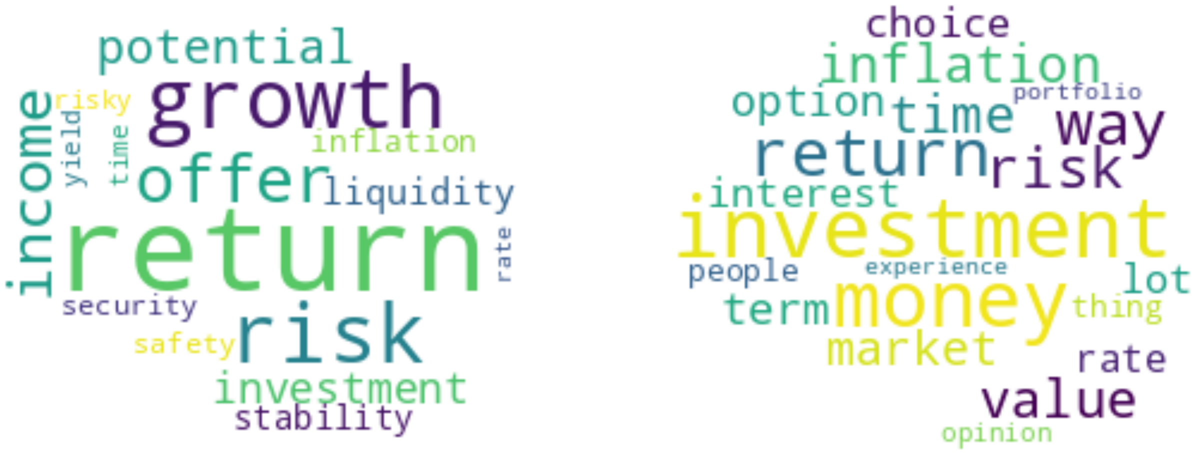
TABLE 2: This table displays the relationship between ratings of the investment options (stocks, bonds, and cash) and demographic characteristics (age, gender, and income). We show the 6 correlations that are significant in both human response data and the GPT4-generated data. A plus sign in the second 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 GPT4 agents both prefer keeping cash more than their younger counterparts.

4 Understanding explanations of ratings

In Section 3, we have shown that the categorical investment ratings generated by GPT4 are closely aligned with human survey ratings across different demographic groups. We now take advantage of the free-form justifications (the “why?" questions accompanying the ratings) to examine the extent to which GPT4 matches the *reasoning* provided by human participants.

4.1 Most common themes

We begin by conducting a simple word count of all nouns that appeared in human and GPT4 responses, separately, in order to determine the most frequent themes being discussed. Table A1 in the 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 GPT4-generated



(a) GPT4 word cloud

(b) Human word cloud

Figure 1: Word clouds of GPT4 and human responses constructed based on the frequency of words in all of the GPT4-generated explanations. We only show words that are used as nouns in the explanation because we are interested in detecting the major themes.

responses (on the right). The human responses are a bit less concentrated than the GPT4-generated responses (the most common term in GPT4-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 1 offers a graphical illustration of the most common terms using word clouds. There are differences in terms of the terms used—human responses focus on “investment" and “money," whereas GPT4 generates more discussion of “return" and “growth", but the general patterns are similar: both sets of explanations concentrate on financial tradeoffs and rewards.

4.2 Discussions of risk and returns in human and GPT4 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 these two semantic dimensions using embeddings.[3] We

[3] 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

begin with the embeddings of four principal sentences:

- Return
  1. "This asset has very high return."
  2. "This asset has very low return."
- Risk
  1. "This asset has very high risk."
  2. "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, as shown in Figure 2, by taking the difference between the vectors $\mathbf{V}^h_{\text{ret}}$ and $\mathbf{V}^l_{\text{ret}}$, we can obtain a vector axis pointing from low to high returns:

$$\mathbf{V}_{\text{ret}} = \mathbf{V}^h_{\text{ret}} - \mathbf{V}^l_{\text{ret}}. \tag{1}$$

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

$$\mathbf{V}_{\text{risk}} = \mathbf{V}^h_{\text{risk}} - \mathbf{V}^l_{\text{risk}}.^{4} \tag{2}$$

Next, we extract the embedding of each explanation in the human response data and the GPT4-generated data. We decompose the meaning of each explanation into three components: return-related, risk-related, and other. Given an embedding vector $\mathbf{emb}_i$ of explanation i, the decomposition can be computed as

$$\mathbf{emb}_i = c_{i,r}\mathbf{V}_{\text{ret}} + c_{i,v}\mathbf{V}_{\text{risk}} + \epsilon_i, \tag{3}$$

build a numerical scale of a meaning (An, Kwak and Ahn, 2018).

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

where $c_{i,r}$ is the projection coefficient of $\mathbf{emb}_i$ onto $\mathbf{V}_{\text{ret}}$, $c_{i,v}$ is the projection coefficient onto $\mathbf{V}_{\text{risk}}$, and ϵ_i is the remaining vector component that does not correspond to either the return vector or the risk vector. Intuitively, $c_{i,r}$ is explanation i's association with high return, and $c_{i,v}$ is i's association with high risk.

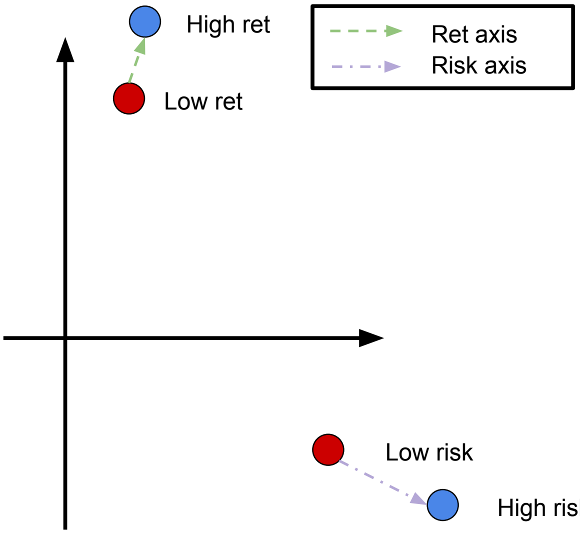


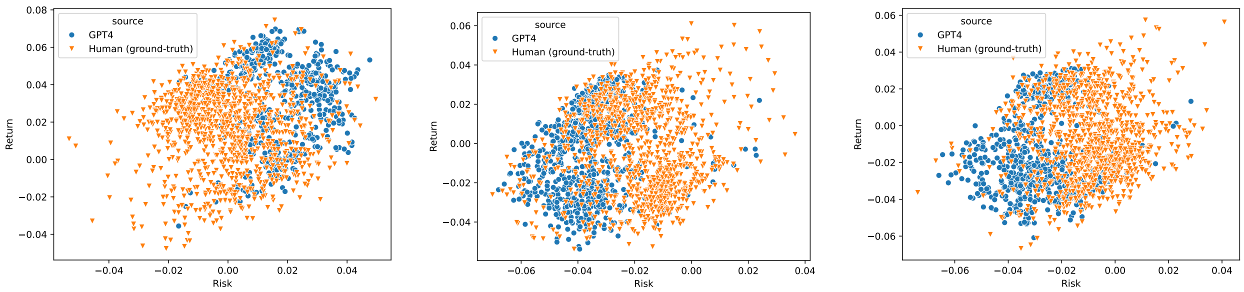
Figure 2: This is a two-dimensional illustration of the risk-return embeddings. The blue dots are the positions corresponding to the embedding vectors of the two sentences describing an asset with high return and high risk, respectively. The two red dots correspond to low return and risk, respectively. The two dashed arrows represent the direction of the return axis and risk axis from low to high. The positions of the dots are only for illustrative purposes. The actual embeddings are 1536-dimensional.

We begin by confirming that GPT4-generated explanations 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' explanations and GPT4-generated explanations have, on average, the highest return components when explaining rankings of stocks, followed by bonds, then cash. Similarly, both human and GPT4-generated explanations project the highest risk when discussing stocks, followed by cash, and then bonds. Overall, GPT4's relative discussion of stocks, bonds, and cash along both return and risk axes is consistent with the verbal explanations provided by human survey participants.[5]

Next, we analyze the differences in the magnitudes of the projected risk and return

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

values across human and GPT4-generated data. Specifically, we study the differences in $\mathbb{E}(c_{i,r})$ and $\mathbb{E}(c_{i,v})$ between human and GPT4-generated explanations to understand whether GPT4 exhibits over- or under-estimation bias in terms of predicting how human participants reason about the risk and return of each investment option.



(a) Stocks

(b) Bonds

(c) Cash

Figure 3: GPT and human's perception of stocks, bonds, and cash with respect to having high return and high risk. Yellow dots are samples from the human survey (ground-truth) data, and blue dots are from the GPT4-generated data. The left subfigure shows humans' and GPT4's perception of stocks, the middle one shows their perception of bonds, and the rightmost subfigure is for cash.

Figure 3 shows that the GPT4-generated explanations of preferences regarding individual investment options tend to have more extreme values of both risk and return compared to human explanations. First, there is a clear vertical and horizontal separation (both return and risk) between the distributions of human and GPT4-generated explanations of ratings of stocks as an investment option. GPT4 agents perceive stocks' return-generating potential as 27.6% higher and their risk as 33.0% higher than human survey participants ($p < 0.001$ for both). This means that GPT4-generated explanations correspond to higher perceptions of stocks' returns *and* higher perceptions of stocks' risks. Similarly, there is a horizontal and vertical separation in the perceptions of keeping cash, in the opposite direction: both humans and GPT4 rank cash as low-risk and low-reward, but GPT4 goes further in both of these negative directions. More specifically, simulated GPT4 agents perceive cash's potential to generate high returns as 15.6% lower and its risk as 29.4% lower than human survey participants ($p < 0.001$ for both). Finally, for bonds there is a clear horizontal separation—both human participants and simulate GPT4 agents rank bonds as low-risk, but GPT4 more strongly so by 35.3%—though a less clear distinction in the return space.

Next, we combine the information from the human and GPT4-generated free-form explanations and the ratings on different investment options to estimate the correlation between the perceived risk and return and the stated rating of each investment option across human participants and GPT4 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 align between actual human survey responses and GPT4-generated responses. In order to address these questions, we estimate the following specification:

$$\begin{aligned}\text{Rating}_{i,k} =& \beta_1 c_{i,k,r} + \beta_2 c_{i,k,v} + \beta_3 \mathbb{1}(\text{i is a GPT4 agent}) \\ &+ \beta_4 \mathbb{1}(\text{i is a GPT4 agent}) * c_{i,r} + \beta_5 \mathbb{1}(\text{i is a GPT4 agent}) * c_{i,r} + \delta_k + \epsilon_{i,k}, \quad (4)\end{aligned}$$

where $\text{Rating}_{i,k}$ is participant i's rating of investment option k (where $k \in \{\text{stocks, bonds, cash}\}$), 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, and β_4 (β_5) can be interpreted as the under- or overstatement of the importance of return (risk) by GPT4.

The results are presented in Table 3, which shows that high perceived return-generating potential of an asset in the free-form responses is significantly positively correlated with higher categorical ratings for that asset. In the human survey, a one-standard-deviation increase in the perceived return is associated with a 0.79 standard deviation increase in the rating. In addition, higher perceived risk in free-form responses is significantly negatively associated with categorical ratings. A one-standard-deviation increase in perceived risk in human survey responses is associated with a 0.56 standard deviation decrease in the associated rating. The GPT4 responses yield a very similar correlation between ratings and returns, with no significant bias. GPT4 responses do show a slightly weaker relationship between risk and returns than human responses; however, the magnitude of the coefficient (0.05) is very small. Overall, the free-form responses capture important

	Dependent variable:
	rating
return	0.788*** (0.015)
generated by GPT4	-0.172^{***} (0.017)
risk	-0.564^{***} (0.017)
GPT4 bias: return	-0.017 (0.019)
GPT4 bias: risk	0.053*** (0.020)
Observations	6,348
R^2	0.554
Adjusted R^2	0.554
Residual Std. Error	0.668 (df = 6340)
Note:	*p<0.1; **p<0.05; ***p<0.01

TABLE 3: Statistical analysis of the correlations between risk and return projections of free-form explanations and the categorical ratings. This table shows the correlations based on human data and the differences in the correlations between the human data and the GPT4-generated data.

information about the reasoning behind the ratings, with high correlations between categorical ratings and the loading on return (positive) and risk (negative) in the free-form responses. Most importantly, we observe that GPT4-generated responses match human responses quite closely in terms of the “reasoning" behind the ratings, with very similar relationships between ratings and risk/return across the actual human survey responses and the GPT4-generated responses.

4.3 Other themes in free-form responses

So far, we have observed that the two major common themes in free-form responses (both, actual responses from survey participants and GPT4-generated responses) are risk and return. We now evaluate which other auxiliary themes play a role in the data, and to which extent these agree between human survey data and GPT4-generated responses. We focus specifically on explanations for stock ratings, to see whether we can speak to the drivers of low stock market participation, which is a long-standing and significant issue documented by the prior literature.[6]

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 (< 3) and explanations associated with positive ratings (> 3). We then use GPT4'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 generate the auxiliary themes based on the GPT4-generated responses. We repeat the procedure three times to ensure that we obtain consistent, replicable summaries and do not capture some noise due to the randomness of GPT4; the themes from each run are listed in Table A3 in the Appendix.

In all summaries, the themes in both the human data and the GPT4-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, GPT4-generated responses reflect the themes “Knowledge and Complexity" (second and third run). Beyond knowledge and understanding, one other theme that appears frequently in both human and GPT4-generated data is “experience" (positive or negative emotional reactions to past

[6] Although our main focus with the auxiliary free-form response analysis is on stocks due to the central role of the low stock market participation issue, for completeness, we provide analogous analysis of the auxiliary themes of investing in bonds (where knowledge is the main theme emerging from the summarization) and cash (where the most consistent theme is the convenience/accessibility of cash) in Appendix D.

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 GPT4-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 and understanding 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)
- 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 GPT4-generated) rating about investing in stocks onto this axis to compute its relevance to knowledge about the stock market. Figure 4 shows a graphical representation of this projection, in blue for the explanations provided by human participants and in yellow for the explanations generated by GPT4.

Both projections center just below zero, with tails in both positive and negative directions. The GPT4 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 explanations and from GPT4-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 $\boldsymbol{\mu} = (\mu_1, \mu_2)$ be the initial means of the two clusters, $\boldsymbol{\Sigma} = (\Sigma_1, \Sigma_2)$ be the initial covariance matrices, and $\boldsymbol{\pi} = (\pi_1, \pi_2)$ be the initial mixing coefficients.

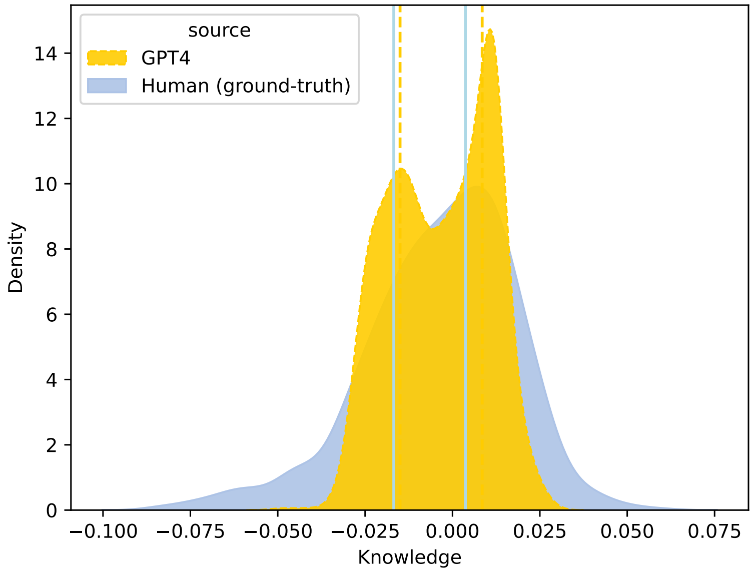


Figure 4: The density of GPT4 agents' responses' and human responses' relevance to having a high amount of knowledge about the stock market. The blue density plot represents the human data, and the yellow plot represents the GPT4 data. The dotted vertical lines represent the center of the three clusters of GPT4 data and the continuous vertical lines are the center of the three clusters of the human data.

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

$$r_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{2} \pi_j \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)},$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate Gaussian distribution.

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.[7]

The resulting clusters are displayed in Figure 4, in solid blue vertical lines for the human survey explanations and dashed yellow vertical lines for the GPT4-generated explanations. The clusters of GPT4-generated explanations are slightly more positive than those of the human survey participants' explanations—cluster centers are slightly further on the right—but the clusters are qualitatively similar. In both cases, cluster 1 (containing explanations of stock ratings that reflect a low amount of knowledge about the stock market) is around -0.015, and cluster 2 (which contains explanations conveying a high level of knowledge about the stock market) is around 0.007. Absence of knowledge and understanding of the stock market is somewhat more prevalent in the GPT4 data than in human responses: in the human distribution, the low knowledge cluster contains 35% of the individuals, and in the GPT4 distribution, the corresponding cluster contains 50% of the agents.

In Table 4, we examine how knowledge levels vary across demographic characteristics, and whether these patterns are correctly reflected by GPT4. 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 human individuals (GPT4 agents) 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. For each data set (embeddings of human responses and embeddings of GPT4-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 men's and higher-income individuals' answers reflect a higher level of knowledge about the stock market. Similar patterns hold in GPT4-generated data: simulated men's 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). GPT4-generated explanations are also more likely to project higher knowledge of the stock market from younger individu-

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

als, although the association between age and knowledge in the actual survey responses is null.

	Dependent variable:	
	Knowledge about the stock market (cluster label)	
	Human	GPT4
age	-0.001	-0.015^{***}
	(0.001)	(0.002)
gender	-0.092^{***}	-0.082^{***}
	(0.025)	(0.027)
income	0.002^{***}	0.058^{***}
	(0.0004)	(0.003)
Observations	1,074	1,042
R^2	0.045	0.241
Adjusted R^2	0.043	0.239
Residual Std. Error	0.403 (df = 1070)	0.436 (df = 1038)
F Statistic	16.893^{***} (df = 3; 1070)	109.795^{***} (df = 3; 1038)
Note:		*p<0.1; **p<0.05; ***p<0.01

TABLE 4: This table shows the correlations between age, gender, and income and the knowledge about the stock market reflected in free-form responses, for human responses in the left column and GPT4-generated responses in the right column. Income is scaled in thousands of dollars.

We conduct similar analyses along the "personal experience" dimension. We use the difference between the embeddings of the following two descriptions to compute a numerical representation of the personal investing experience dimension:

- I have had very good experiences investing in the stock market. (positive experience)
- I have had terrible experiences investing in the stock market. (negative experience)

Figure 5 shows the resulting projections, and we apply the same approach as we did for the knowledge theme (i.e., a Gaussian Mixture Model) to cluster these distributions, defining three clusters corresponding to more positive experiences, more negative experiences, and neutral experiences. The solid blue vertical lines in Figure 5 mark the

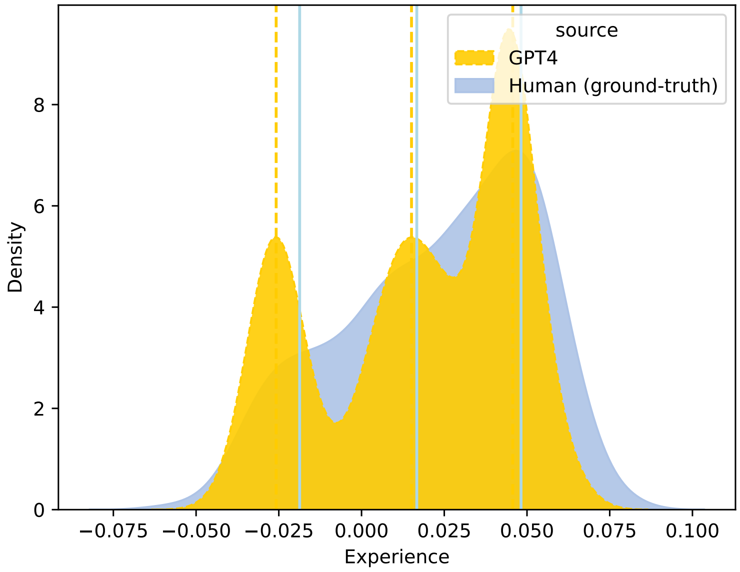


Figure 5: The density of GPT4 agents' responses' and human responses' relevance to positive experience about the stock market. The blue density plot represents the human data, and the yellow plot represents the GPT4-generated data. The continuous vertical lines mark the centers of the three clusters of human responses, and the dotted vertical lines mark the centers of the three clusters of GPT4-generated responses.

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 GPT4-generated responses.

In each data set, there is one cluster with responses reflecting negative past experiences about the stock market, centered around -0.023, one with mildly positive experiences, centered around 0.018, and one with very positive experiences, centered around 0.049. Moreover, the distributions across the three clusters are very similar between human responses and GPT4-generated responses. 23% of human participants' explanations (24% of GPT4-generated explanations) are in the cluster associated with negative experiences with the stock market, 33% of human explanations (35% of GPT4-generated explanations) are in the cluster with slightly positive experiences with the stock market, and 44% of GPT4 agents (41% of human participants) are in the cluster with the most positive experiences with the stock market.

We examine how reported personal experiences with the stock market vary across demographics, and whether GPT4 is able to capture those differences. We regress the cluster

	Dependent variable:	
	Experience with the stock market (cluster label)	
	Human	GPT4
age	-0.004^{**}	-0.027^{***}
	(0.002)	(0.003)
gender	-0.201^{***}	-0.093^{**}
	(0.047)	(0.041)
income	0.003^{***}	0.110^{***}
	(0.001)	(0.005)
Observations	1,074	1,042
R^2	0.037	0.335
Adjusted R^2	0.035	0.333
Residual Std. Error	0.774 (df = 1070)	0.648 (df = 1038)
F Statistic	13.866^{***} (df = 3; 1070)	174.154^{***} (df = 3; 1038)
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

TABLE 5: This table shows the correlations between age, gender, and income and human participants' (GPT4 agents') experience with the stock market. Income is scaled in thousands of dollars.

labels (1 for the negative experience cluster, 2 for mildly positive experiences, and 3 for strongly positive experiences) on age, gender, and income of the corresponding human participant or simulated GPT4 agent. Table 5 reports 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 exactly the same in GPT4 data: simulated younger individuals, men, and individuals with higher incomes generate explanations that project more positive experiences with the stock market.

Together, Tables 4 and 5 show that there are strong demographic patterns in the expressions of both knowledge and personal experiences of the stock market and that GPT4 does a good job capturing these directional patterns across demographics.

5 Are humans and GPT4 transitive?

A necessary condition for economic preferences to be considered rational is transitivity. For example, if a rational consumer prefers stocks over bonds and bonds over cash, then he or she must also prefer stocks over cash. In this section, we examine the transitivity property of human and GPT4-generated preference orderings. For this analysis, we leverage the relative-comparison responses, where participants were asked to directly order pairs of investment options within each question. For example, when we elicit preferences between stocks and bonds, we allow participants to choose one of three responses: stocks, bonds, or indifferent.

To test whether each individual's preference 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:*

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

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

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

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

Applying proposition 1, we find that overall 84.4% of human survey responses and 98.7% of GPT4-generated responses follow transitivity. This means that GPT4 agents' preferences are almost always transitive, while human responses may not be. We further investigate the cause of this discrepancy between humans and simulated GPT4 agents. In particular, we study the source of intransitivity in the human data set.

We start by dividing the human sample into responses from men and responses from women. As shown in Table 6, we observe that within the male subsample, 89.5% of the participants have transitive preferences, and within the female subsample, only 79.1% of the respondents satisfy transitivity. Furthermore, we observe that conditioning on human participants with no "indifferent" responses, 95.7% of these participants have transitive preferences. However, conditioning on participants with at least one "indifferent," only 72.0% of human participants have transitive preferences.[8]

	GPT4		Human			
n obs	1042		1074			
transitive prob	98.7%		84.4%			
difference	14.3%***					
	Male	Female	Male		Female	
n obs	500	542	544		530	
transitive prob	99.8%	97.6%	89.5%		79.1%	
difference	2.2%***		10.4%***			
			diff	**indiff**	**diff**	**indiff**
n obs			325	219	235	295
transitive prob			96.6%	79.0%	94.5%	66.8%
difference			17.6%***		27.7%***	

TABLE 6: Preference transitivity difference of human participants and simulated GPT4 agents. In this table, the "diff" columns report the results conditioned 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 preferences is due to an imbalanced distribution of participants who respond with at least one "indifferent." Indeed, we find that 40.3% of male human participants with at least one indifference, while the proportion for female participants with at least one indifference is 55.7%. In addition, we observe that conditioning on the human participants did not respond with at least one indifference, males are still more likely to have transitive preferences than females: 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%). In both genders, this share is lower than the share of simulated GPT4 agents with transitive preference.

Overall, we make four observations related to rationality of expressed preference orderings:

[8] Note that no simulated GPT4 agent responded with "indifferent" despite being provided this option. This is potentially because GPT4 is trained to produce answers that are preferred by human requesters, and when a human asks GPT4 to make a selection, "indifferent" is usually not a desirable answer.

1. GPT4 agents almost always have transitive preference orderings over investing in stocks, bonds, and cash, but human survey participants are less likely to have transitive preference orderings.

2. The main driver of non-transitive preferences in human survey responders is the presence of at least one indicated indifference between investment options.

3. There is a gender difference in transitive preferences, whereby men have display more transitive orderings than women. This is driven by both the frequency of indifference (higher in women than in men) and the likelihood of violating transitivity conditional on having at least one indifferent response (33.2% for women vs. 21.0% for men).

4. Even among male human participants with no indifferent responses (the group of huan participants with the highest share of transitive preference orderings), the share of transitive preference orderings is still lower than that of simulated GPT4 agents, but this difference is smaller in both size and statistical significance (96.6% vs. 99.8%).

6 Conclusion

We examine how well generative AI (as exemplified by OpenAI's GPT4) can replicate human investment preferences, especially across demographics. Algorithmic bias is becoming an increasingly important issue with the growing use of machine learning in finance. For example, Bartlett et al. (2022) showcase the issues in the credit space, and multiple banks, including Wells Fargo, have been sued for using AI models that led to discriminatory lending. Our results show a more positive side of AI: in the context of predicting investment preferences, generative AI does not seem to suffer from systematic bias and correctly captures heterogeneity across gender, age, and income. Not only are GPT4-generated ratings of investment options highly correlated with actual survey participants' ratings, but GPT4 also has similar "reasoning"—capturing the main themes in free-form explanations.

These results have important implications for the financial services industry, especially the rising prominence of robo-advising in the investment management space (D'Acunto, Prabhala and Rossi, 2019; Rossi and Utkus, 2021). Our analysis suggests that the recent advances in generative AI will enable further growth for these types of automated investment advice services, correctly capturing heterogeneous preferences across investors while avoiding some of the pitfalls with more "rational" allocations. This brings two advantages. First, our results showcase the potential of generative AI to improve and streamline the tailoring of algorithmic financial advice, reflecting the reasoning and preferences of specific demographic groups. Second, while GPT4-generated responses are highly correlated with human survey responses, there is one dimension on which they consistently outperform: while human responses can violate the transitivity axiom, especially when expressing indifference, GPT4-generated responses are practically always transitive while reflecting the same general patterns and reasoning.

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.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson. 2024. "Artificial intelligence, firm growth, and product innovation." *Journal of Financial Economics*, 151: 103745.

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*.

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

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender shades: Intersectional accuracy disparities in commercial gender classification." 77–91, PMLR.

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

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.

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.” *The Journal of Finance*, 59(1): 137–163.

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

Li, Edward Xuejun, Zhiyuan Tu, and Dexin Zhou. 2023. “The Promise and Peril of Generative AI: Evidence from ChatGPT as Sell-Side Analysts.” *Available at SSRN 4480947*.

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

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*.

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

Veldkamp, Laura. 2023. "Valuing data as an asset." *Review of Finance*, 27(5): 1545–1562.

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

Zou, James, and Londa Schiebinger. 2018. "AI can be sexist and racist—it's time to make it fair."

Appendix

A Additional Tables

15 most frequent nouns in human responses		15 most frequent nouns in GPT4 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 GPT4 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
GPT4 Batch 1				
Volatility and Predictability	Income and Risk Tolerance	Perception of the Stock Market	Risk versus Reward	Emotional and Psychological Responses
GPT4 Batch 2				
Perception of Volatility	Income Level Concerns	Market Predictability	Time Horizon	Knowledge and Complexity
GPT4 Batch 3				
Volatility and Stability	Investment Horizon	Income Considerations	Knowledge and Complexity	Emotional Response and Comfort Level

TABLE A3: Additional themes from GPT-4 and Human Responses. We collected 3 batches of 5 themes each by using GPT4 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.

B Experimental Instructions

B.1 Human Survey Instructions

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 the 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 C1—appear sequentially, in random order.]

[Questions with comparisons of stocks versus bonds, stocks versus cash, and bonds versus cash—following the example in Figure C2—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?

B.2 Sample GPT4 Prompt

> *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?*

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.

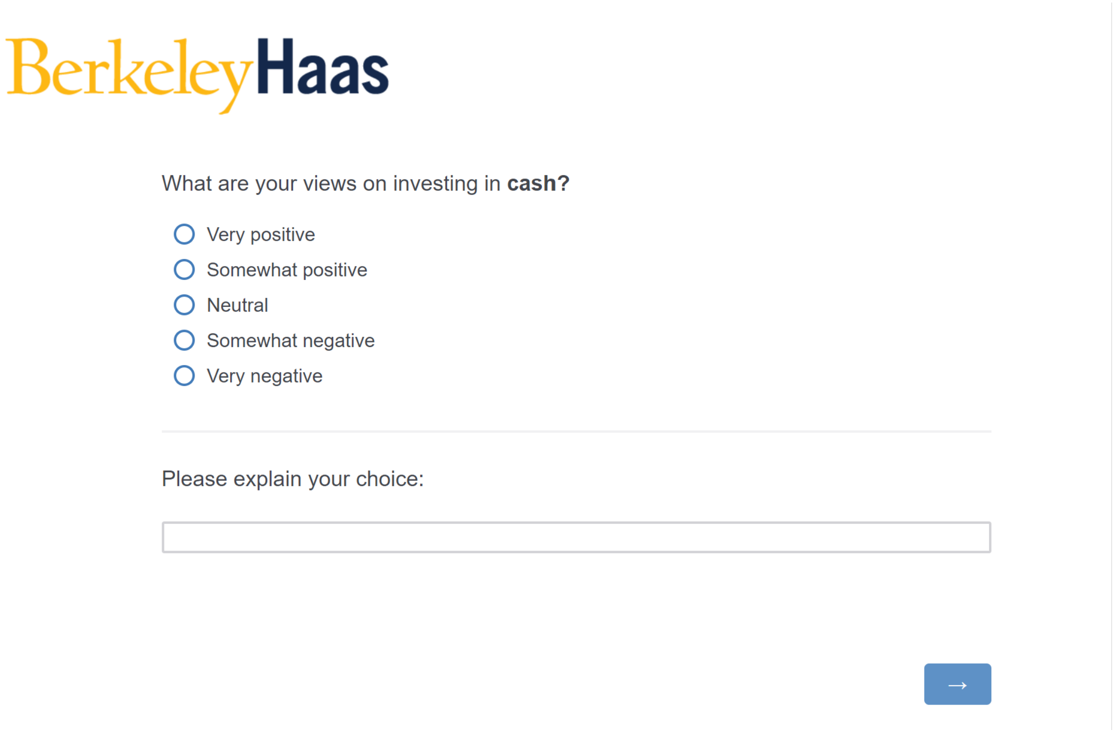
Answer to question 3: neutral. Explanation: Cash has no growth potential.

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

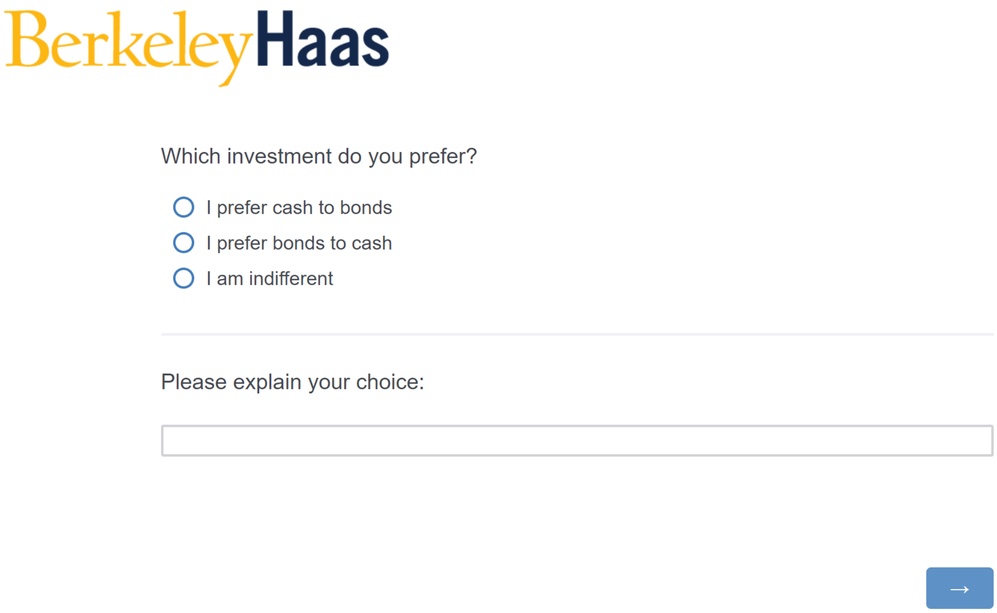


What are your views on investing in **cash?**

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

Please explain your choice:

Figure C1: Screenshot of a single-rating question from the human survey.



Which investment do you prefer?

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

Please explain your choice:

Figure C2: Screenshot of a relative-comparison question from the human survey.

C Supplemental discussions

C.1 Proof of proposition 1

Proposition:

Let $a \in \{\text{stocks}, \text{bonds}, \text{indifferent}\}$, $b \in \{\text{stocks}, \text{cash}, \text{indifferent}\}$, and $c \in \{\text{cash}, \text{bonds}, \text{indifferent}$ 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:

1. If $a, b, c \neq$ indifferent, exactly one of the following must be true $a = b$ or $b = c$ or $a = c$.
2. If $a =$ indifferent, then either $b = c$ or $b, c \in \{\text{stocks}, \text{bonds}\}$.
3. If $b =$ indifferent, then either $a = c$ or $a, c \in \{\text{cash}, \text{stocks}\}$.
4. If $c =$ indifferent, then either $a = b$ or $a, b \in \{\text{cash}, \text{bonds}\}$.

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, assume there is only one indifference in the preference relation, the two options that are indifferent must either both 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).

C.2 Introduction to neural networks

We start by discussing a simple neural network: a one-hidden-layer linear neural network.[9] This network is defined by three dimensions: input dimension $\dim_{in}$, hidden dimension $\dim_h$, and output dimension $\dim_{out}$.

There are two mappings in this one hidden layer network. The first is the mapping from the input space to the hidden space. Mathematically, let the input data be X

$$H = f_1(X) = XW_1 + B_1,$$

. Then the hidden space H can be represented as where W_1 and B_1 are trainable matrices

[9] This introduction of neural networks and the self-attention mechanism are based on the discussion in Hochberg et al. (2023).

of parameters.

The second function maps from the hidden space H to the output space Y:

$$Y = f_2(H) = f_2(f_1(X)) = (XW_1 + B_1)W_2 + B_2,$$

$$= XW_1W_2 + B_1W_2 + B_2,$$

where W_2 and B_2 are trainable matrices of parameters.

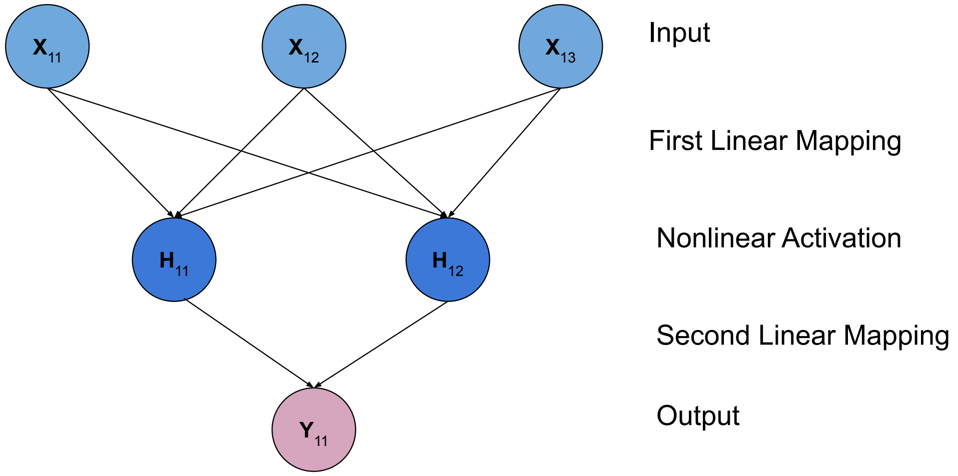


Figure B1: Diagram of a one-hidden-layer neural network with a nonlinear activation function. The top blue circles are the inputs, which, in the case of this figure, are three-dimensional vectors. The green circles in the middle are the nodes in the hidden layer, and the violet circle on the bottom is the output.

As shown in Figure B1, this simple linear network can be generalized by adding a nonlinear activation function $g(\cdot)$. When we apply the activation function to the first mapping, the output of the first mapping (the hidden space) becomes

$$H = g(f_1(X)) = g(XW_1 + B_1).$$

Similarly, we can apply a nonlinear activation function after any mapping to add nonlinearity to a linear transformation, allowing neural networks to more flexibly fit any relation between the input X and output Y.

C.3 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, then we introduce the self-attention mechanism which allows GPT4 to be aware of the 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.

C.3.1 Decoder

The original Transformer architecture was designed for tasks like 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.

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

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 incorporates 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.

C.3.2 Self-attention

The goal of self-attention is to create a numerical embedding for each piece of text, and this embedding is created to respect 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, this text is broken into sub-word tokens in a parsing process called tokenization.

This set of tokens are pre-defined 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, suh as “happy.” Many other tokens capture short and common words like “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.

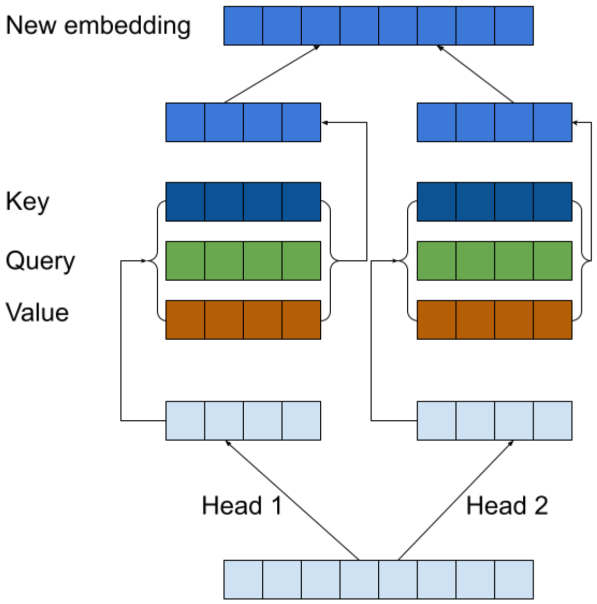


Figure B2: Diagram of a multi-head attention layer

As shown in Figure B2, 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. Then, 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 multiple equal-sized sub-vectors. 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.

C.3.3 Pre-training

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

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

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

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

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

C.3.4 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.

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.

C.4 Definition of correlation coefficients

- **Pearson Correlation Coefficient**

  Formula:

  $$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

  The Pearson correlation coefficient measures the linear relationship between two continuous variables. The formula calculates the covariance of the variables normalized by the product of their standard deviations. The correlation coefficient r_{xy} ranges from -1 to 1, where $r = 1$ indicates a perfect positive correlation, $r = -1$ indicates a perfect negative correlation, and $r = 0$ indicates no correlation. When r is close to 1 or -1, it suggests a strong correlation between the variables.

- **Spearman's Correlation Coefficient**

  Formula:

  $$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

  Spearman's rank correlation coefficient assesses the monotonicity of the relationship between two variables. The formula computes the differences between the ranks of corresponding data points, squares them, sums them up, and normalizes the result. The coefficient ρ ranges from -1 to 1, where $\rho = 1$ indicates a perfect positive correlation, $\rho = -1$ indicates a perfect negative correlation, and $\rho = 0$ indicates no correlation.

- **Kendall's Correlation Coefficient**

  Formula:

  $$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}.$$

Similar to Spearman's coefficient, Kendall's correlation coefficient also evaluates the ordinal (rank) association between two variables. It counts the number of concordant and discordant pairs of observations and normalizes them. A concordant pair refers to a pair of observations where the ranks are in the same order for both variables. That is, if in the first variable, X has a higher rank than Y, and in the second variable, X also has a higher rank than Y, then this pair is considered concordant. A discordant pair is the opposite. The coefficient τ ranges from -1 to 1, where $\tau = 1$ indicates perfect agreement between the rankings, $\tau = -1$ indicates perfect disagreement between the rankings, and $\tau = 0$ indicates no association between the rankings.

D Analysis of explanations of cash and bonds ratings

In this section, we use GPT4 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 GPT4'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.

D.1 Investing in bonds

The major common theme discovered in explanations bond ratings is the level of "knowledge" and unnderstanding about financial markets (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.
- I do not know anything about the bond market.

As shown in Figure B3, the distributions of the embeddings of human and GPT4-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 -0.05 and -0.02) and one mildly positive cluster (centered at 0.01). The GPT4 distribution contains one negative cluster (centered at -0.02), one neutral cluster (centered at -0.004), and one positive cluster (centered at 0.01).

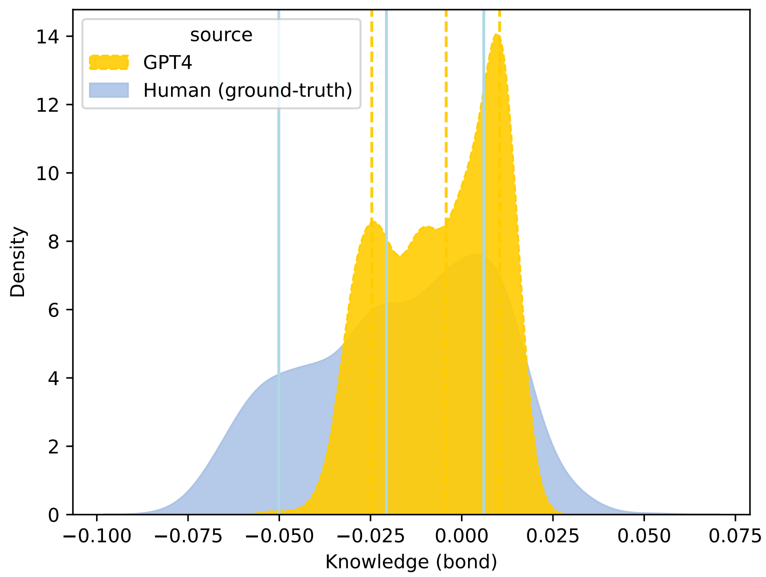


Figure B3: The density of human and GPT4-generated responses' relevance to having a high amount of knowledge about the bond market. The blue density plot represents the human data, and the yellow plot represents the GPT4-generated data. The continuous vertical lines mark the centers of the three clusters of human responses, and the dotted vertical lines represent the centers of the three clusters of GPT4-generated data.

Table A4 shows the demographic variations associated with differences in agents' knowledge about the bond market. Older, male, and higher-income human participants tend to be more knowledgeable about the bond market (significant at the 5% level). Similarly, older and higher-income GPT4 agents have a better understanding of the bond market. The one disagreement between GPT4-generated data and actual human survey data is the role of gender: while human men expess significantly more knowledge about the bond market than women, simulated GPT4 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).

D.2 Keeping cash

Next, we analyze the themes of the explanations corresponding to positive versus negative attitudes towards holding cash. We find that the most consistent theme in these responses, outside of risk and return, is "accessibility," or the convenience benefits of holding cash, including its ready liquidity. Some individuals rate cash highly because money is readily accessible at any time. We construct the accessibility dimension using

	Dependent variable:	
	Knowledge about the bond market (cluster label)	
	Human	GPT4
age	0.004^{**}	0.011^{***}
	(0.002)	(0.004)
gender	-0.269^{***}	0.087^{*}
	(0.048)	(0.048)
income	0.003^{***}	0.041^{***}
	(0.001)	(0.006)
Observations	1,074	1,042
R^2	0.054	0.058
Adjusted R^2	0.051	0.055
Residual Std. Error	0.791 (df = 1070)	0.768 (df = 1038)
F Statistic	20.391^{***} (df = 3; 1070)	21.118^{***} (df = 3; 1038)
Note:		*p<0.1; **p<0.05; ***p<0.01

TABLE A4: This table shows the correlations between age, gender, and income and human participants' (GPT4 agents') knowledge about the bond market. Income is scaled in thousands of dollars.

the following sentences:

- I like the high level of accessibility of cash.
- I do not care about the level of accessibility of cash.

As shown in figure B4, the distributions of the embeddings of human and GPT4 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 0.04) and one nearly neutral cluster (centered at 0.0002 and containing 58% of the participants). The GPT4 distribution also contains one positive cluster (centered at 0.05) and one nearly neutral cluster (centered at 0.02 and containing 79% of the simulated GPT4 agents).

Table A5 shows the demographic variations associated with differences in agents' level of consideration of the high accessibility of cash. Older and lower-income GPT4 agents care more about the high accessibility of cash. However, human participants do

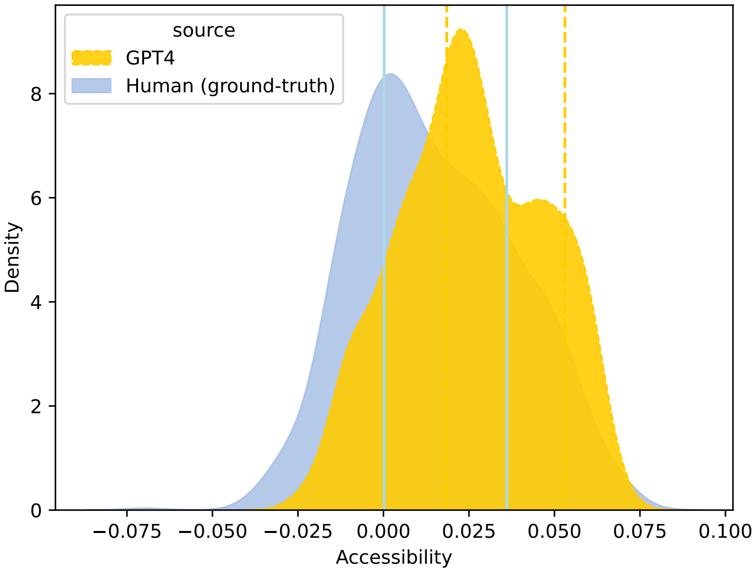


Figure B4: The density of the relevance of human and GPT4-generated explanations of cash ratings to caring about the high accessibility of cash. The blue density plot represents the human data, and the yellow plot represents the GPT4-generated data. The continuous vertical lines mark the centers of the two clusters of human responses, and the dotted vertical lines represent the centers of the two clusters of GPT4-generated data.

not show strong demographic differences in terms of the importance assigned to accessibility of cash.

	Dependent variable:	
	Caring about the accessibility of cash (cluster label)	
	Human	GPT4
age	0.001	0.007^{***}
	(0.001)	(0.002)
gender	−0.010	0.039
	(0.030)	(0.026)
income	−0.00004	-0.022^{***}
	(0.0004)	(0.003)
Observations	1,074	1,042
R^2	0.001	0.056
Adjusted R^2	−0.002	0.053
Residual Std. Error	0.492 (df = 1070)	0.406 (df = 1038)
F Statistic	0.432 (df = 3; 1070)	20.364^{***} (df = 3; 1038)
Note:		*p<0.1; **p<0.05; ***p<0.01

TABLE A5: This table shows the correlations between age, gender, and income and human participants' (GPT4 agents') level of consideration of the high accessibility of cash. Income is scaled in thousands of dollars.