Is Artificial Intelligence (AI) Risk-Averse?

Yuree Lim*

Abstract

This study evaluates the efficacy of ChatGPT-generated data in replicating human survey responses, focusing on risk aversion (RA) metrics. Our analysis finds that while ChatGPT can replicate overall variations and capture marginal associations, it introduces biases and is sensitive to specific personas. Secondly, we document significant behavioral biases in Generative AI agents, particularly bias toward risk-seeking behavior. Finally, we reveal that ChatGPT exhibits heightened risk aversion for specific demographic groups, including females, older individuals, those with lower income, less education, and unemployment. Our results highlight the usage of AI-generated data in research, reinforcing the potential behavioral bias of AI in simulating human-like decision-making patterns.

Keywords: artificial intelligence, risk aversion, large language models

JEL: G10, G11, G12

* Merrilee Alexander Kick College of Business & Entrepreneurship, Texas Woman's University, ylim2@twu.edu

Is Artificial Intelligence (AI) Risk-Averse?

1. Introduction

The rapid advancement of artificial intelligence has opened up new possibilities for data generation and analysis. Eisfeldt and Schubert (2024) highlight the need for adding Generative AI to the research toolbox. One such innovation is the ability of language models, like ChatGPT, to generate data that mimics real human responses. This paper explores the efficacy of data generated by ChatGPT in replicating human survey responses, specifically focusing on risk aversion (RA) metrics.

At least one start-up in private industry suggests that "synthetic users" can supplement or replace human respondents in development and marketing,[1] and there is a growing interest among polling companies to explore the opportunities promised by synthetic data.[2] But can a pretrained LLM produce synthetic responses for respondent personas that accurately mirror what similar human respondents would say on a real survey? In this paper, we evaluate the quality of synthetic data through a series of comprehensive tests. Our investigation begins with univariate tests, comparing correlations, means, and variances between human and GPT-generated RA to assess whether ChatGPT can accurately recover overall variations. We find that while ChatGPT demonstrates a reasonable ability to replicate overall variations, it often introduces biases and exhibits sensitivity to the specific personas represented in the prompts. The correlation analysis reveals a positive relationship between human and GPT RA with differences in risk aversion levels.

Our second approach examines ChatGPT's capability to capture marginal associations between covariates. By comparing regression results derived from both human and synthetic samples, we observe that the patterns of relationships often align, although systematic deviations occasionally emerge. This comparison highlights the potential of synthetic data to approximate human survey results, particularly in exploratory analyses where the signs and statistical significance of associations are critical.

---

[1] https://www.syntheticusers.com.

[2] For example, the American Association for Public Opinion Research has hosted several events discussing the opportunities and challenges in using AI in survey research, such as the NYAAPOR events of May 3, 2023 and October 18, 2023.

Third, we delve into the reasoning processes of ChatGPT. To validate these synthetic values, we prompted ChatGPT to provide explanations for its RA level choices. We observed a surprising consistency in these explanations, which allowed us to correlate keywords with specific RA levels. This analysis shows that keywords associated with high RA levels (e.g., "age," "wealth," "income," "ill," and "financial literacy") are sensible and provide reassurance of the validity of the numeric RA levels. Further validation involved embedding the explanations in semantic space using a BERT transformer. We measured the cosine similarity between explanations corresponding to different RA levels, confirming that explanations for closer RA levels are more similar, while those for distant RA levels are more distinct. This consistency was observed in the HILDA and SCF surveys but less so in the GSOEP survey, indicating potential variability in the model's performance across different datasets. We also employed ChatGPT to annotate a random sample of SCF explanations, comparing the consistency of GPT-generated RA levels. The high correlation between these annotations and the initial GPT-generated RA levels suggests a strong association between the model's numeric outputs and its reasoning processes. Lastly, we used Latent Dirichlet Allocation (LDA) to generate topics from the explanations and calculated the average RA score per topic. This analysis demonstrates a reasonable association between topics and RA levels, further validating the synthetic data.

Fourth, we explore the confidence levels of synthetic risk aversion (RA) data generated by ChatGPT and examine how these confidence measures relate to the variability observed in the data. While ChatGPT does not allow users to directly observe the posterior probabilities of its outputs, we attempted to gauge the AI's confidence in its RA level predictions. The responses varied, but most adhered to a 100%-item scale ranging from not at all confident (0%) to very confident (100%). Our analysis of these self-reported confidence levels reveals that the majority of responses indicate a high degree of confidence, with most values above 70%. We further investigate the AI's confidence across different dimensions, including survey and gender. The results show that ChatGPT's confidence varies by survey, with the GSOEP survey eliciting lower confidence levels compared to the SCF and HILDA surveys. This variation is particularly pronounced for female personas, suggesting that less detailed persona sketches lead to lower confidence in the AI's predictions. Additionally, we examine the relationship between RA levels and confidence. As expected, neutral RA levels are associated with weaker confidence, while higher or lower RA levels correspond to greater confidence, forming a U-shaped pattern. To understand the empirical

variability of the synthetic data, we compare the AI's self-reported confidence with the standard deviation of RA levels. The analysis indicates a negative relationship: higher confidence corresponds to lower empirical uncertainty. This finding suggests that ChatGPT's self-reported confidence is a reasonable proxy for the empirical variability of its outputs.

Fifth, we address the uncertainty of AI-generated synthetic data by examining the standard deviation of risk aversion (RA) levels produced by ChatGPT because one of the critical challenges in using AI-generated data is measuring the uncertainty associated with such data. Our findings reveal that ChatGPT generates less variation compared to real human data. This reduced variation might be attributed to the algorithm's difficulty in predicting less frequently occurring groups, which constitute a smaller portion of the training data. We start by analyzing the standard deviation at the human-by-survey level, combining demographics such as gender, age, income, education, and employment. Contrary to expectations, our analysis shows little evidence that ChatGPT is more confident for groups that are more commonly represented in the data. The data suggest marginally larger standard deviations for males, except for the SCF survey, where the most commonly appearing group's synthetic standard deviation is slightly lower than that of the least common group. This precision is significantly greater than what is observed among real humans. Furthermore, we explore whether these patterns reflect sample size effects or demographic stereotypes, such as the notion that older females are more risk-averse. While genuine variation in RA levels among specific human groups drives observed differences, our findings indicate that demographic descriptions might influence the synthetic data's variation. Importantly, we observe that the LLM's estimates are consistently more precise than human data across almost all profiles, regardless of sample size. We also provide descriptive plots of the average per-persona standard deviation, broken down by survey and covariates, highlighting the variability of synthetic data across different surveys. Despite some differences, the synthetic data generally exhibit less noise than the human data.

Sixth, we examine the distributions of synthetic RA levels. By aggregating synthetic respondents per human, we plot the distributions by profile and survey, revealing consistent estimates of variance with RA ranges typically spanning between 2 and 4. The distributions illustrate the synthetic data's reliability, especially for certain demographic groups such as older females with no college education.

Lastly, we discuss our research design choices, including the model's "temperature" setting and the selection of AI agents. In the context of ChatGPT, "temperature" refers to the determinism in selecting subsequent tokens based on the given prompt and preceding tokens. A temperature setting of zero ensures the most likely subsequent token is always chosen, while higher temperatures introduce variability, often referred to as "creativity." This study primarily utilized ChatGPT at its most creative setting, with a temperature of 1. While this setting should theoretically produce the noisiest and most representative results of actual human randomness, our findings indicate that even at this high temperature, ChatGPT's estimates were more precise than those of human respondents, with a lower average standard deviation (0.72 vs. 0.77). We observed that the standard deviation in ChatGPT's responses combines two sources of variation: the inherent randomness of the data generation process and the variation stemming from averaging across different demographic groups. To further understand the impact of temperature on the model's performance, we calculated the standard deviation for each profile at various temperature settings (0 to 1). Our analysis highlights that more detailed demographic inputs decrease the average standard deviation of RA levels, with a temperature of 1 yielding the highest standard deviation and the highest proportion of correctly generated RA levels. Consequently, we adopt a temperature setting of 1 for our baseline results. In addition to temperature, we explored the performance of various AI agents, including OpenAI's GPT-4o, Google's Gemini 2.0 Flash (experimental), and Meta's Llama 3.3 70B (versatile) models. Our analysis reveals that GPT consistently provided superior results across all temperature settings, with the best performance at a temperature of 1 (49% accuracy). Despite the slight performance improvement of GPT-4o over GPT-4o-mini, the substantial cost difference between the models led us to use the more cost-efficient GPT-4o-mini, which achieved 41% accuracy at a temperature of 1.

While synthetic data from ChatGPT shows promise for preliminary research and hypothesis testing, several limitations must be considered. The model's tendency to overfit certain demographic profiles and the inherent biases in the training data can lead to discrepancies between synthetic and real data. Furthermore, the model's overconfidence in its predictions necessitates careful interpretation of results, particularly in high-stakes decision-making contexts.

Our research makes significant contributions to the existing literature. Firstly, this study enriches the growing body of knowledge on AI-generated data by offering a detailed evaluation of the reliability and limitations of data produced by AI. By understanding these aspects, we can

better leverage AI to complement traditional data collection methods, ultimately enhancing the efficiency and scope of empirical research. Secondly, our research addresses the behavioral biases of Generative AI agents by documenting a notable bias toward risk-seeking behavior. Finally, our research contributes to the literature on the relationship between demographic information and risk aversion levels by demonstrating that GPT exhibits heightened risk aversion for certain demographic groups, including females, older individuals, those with lower income, less education, and unemployment.

The paper proceeds as follows. Section 2 provides a literature review and discusses our contributions. Section 3 describes the data sources and methodology used in the analysis. Section 4 conducts the main tests. Sections 5 and 6 present the results of reasoning and confidence. Section 7 discusses our research design choices. Section 8 concludes the paper.

2. Literature Review and Contributions

Our research makes significant contributions to several key areas. Firstly, we enhance the literature on the application of Generative AI in financial research by using large language models (LLMs) as tools for simulating survey responses. These LLMs offer a quick, cost-effective, and readily available alternative to human respondents, making them valuable for preliminary survey testing, product or user experience testing, and scenarios where human subjects are unavailable. For instance, Hewitt et al. (2024) demonstrate that GPT-4 can outperform human experts in predicting experimental outcomes. LLMs can also generate tailored financial advice based on the demographics of a human counterpart or assume interviewer roles in data collection. The utility of using an LLM instead of a human subject depends on its ability to match human responses accurately and without bias, the replicability and robustness of the LLM-based method, and the cost-effectiveness compared to human surveys. Bybee (2023) shows that LLM-generated expectations for macroeconomic time series and stock returns closely track real-world survey expectations, and replicate behavioral biases seen in human respondents. Similarly, Fedyk et al. (2024) find that LLMs mimic human preferences over asset classes when prompted with specific demographic characteristics, though LLM responses tend to exhibit more transitive preferences. Furthermore, LLMs can simulate actions for heterogeneous agents in various scenarios, enabling researchers to generate hypotheses on how observed outcomes may arise from complex strategic interactions, which can then be tested in real-world settings (Tranchero et al., 2024). Horton (2023)

finds that LLMs, given specific preferences, can exhibit similar behaviors to humans in simulations, making them useful for piloting studies before real-world testing. Based on the literature, we hypothesize that LLM-generated risk aversion resembles the behavioral biases exhibited by human respondents. Our research suggests that LLMs can provide insights and enhance research methodologies across diverse applications, from financial advice to behavioral studies.

Our research methodology draws on established literature that utilizes prompts and LLM responses to study AI behaviors. For instance, Lo and Ross (2024) asked ChatGPT 4.0 whether one should invest in the stock market, and while it didn't provide real-time advice, it offered general financial guidance such as conducting personal financial assessments, diversifying investments, and considering dollar-cost averaging. Similar to this paper, our paper uses an LLM to gather similar responses. Chang et al. (2024) employed an LLM to assign sentiment scores to earnings calls, while Jha et al. (2024) asked an LLM to evaluate whether earnings calls indicated significant changes in a firm's capital spending. We adopt this approach to extract numerical distinctions from an LLM. Bisbee et al. (2024) show a political bias of LLMs. Moreover, Eisfeldt et al. (2023) utilized LLMs to explain task classifications, providing insights into how generative AI applies rubrics. Our study similarly obtains explanations from an LLM to understand its classifications.

We extend the literature on AI behavioral bias by examining LLM-generated risk aversion. Numerous studies have explored the alignment between LLM and human behaviors, demonstrating that LLMs can exhibit economic behaviors and preferences similar to humans (e.g., Brand et al., 2023; Dillion and Tandon, 2023). Jia et al. (2024) compared various models in terms of loss aversion and risk preferences, highlighting the need for systematic examination of AI-generated risk aversion using large human surveys. [3] Previous research has also evaluated LLM rationality in financial decision-making, with LLMs often scoring higher than humans (Chen et al., 2023). However, these frameworks presuppose that LLMs align with human decision-making processes. Fairness and bias in LLM processing are critical issues, particularly concerning minority groups. Ranjan et al. (2024) highlighted significant biases and performance declines

[3] In addition to household survey, there are other ways of risk elicitation and methodological considerations (Holt and Laury, 2002; Crosetto and Filippin, 2016; Falk et al., 2023). In this paper we focus on systemically available comprehensive survey data.

when LLMs handled information about minority groups. This underscores the need for frameworks that quantitatively evaluate LLM behavior and address biases related to human demographics.

Our study focuses on risk aversion, a concept extensively covered in the literature. Factors such as income, age, gender, and education influence financial risk tolerance (Campbell, 2006; Giannetti and Wang, 2016). For example, Morin and Suarez (1983) found that risk aversion increased with age for low-net-worth households but decreased for high-net-worth households. Other research indicates that female investors tend to be more conservative (Johnson and Powell, 1994), and lower financial literacy correlates with less stock investment (van Rooij et al., 2011). We build on this literature to assess whether LLM-generated risk aversion aligns with these established patterns.

3. Data and Methodology

In this section, we outline our data sources, research design, and methodology for collecting and comparing human risk preferences across different demographics with analogous responses generated by GPT-4 for our benchmarking study.[4]

3.1. Data: Human Surveys

We utilize three datasets for this study: the Survey of Consumer Finances (SCF) for the United States, the Household, Income and Labour Dynamics in Australia (HILDA) survey, and the German Socio-Economic Panel (GSOEP). The SCF data is sourced directly from its official website, while the HILDA and GSOEP datasets are obtained through the work of Gu, Peng, and Zhang (2024).

A first widely used dataset is the U.S. SCF, which provides comprehensive information on households' asset allocations and the risk preferences of all household members. Each respondent answers the following question: which of the following statements comes closest to describing the amount of financial risk that you are willing to take when you save or make investments? The answer options are (1) I take substantial financial risks expecting to earn substantial returns; (2) I take above-average financial risks expecting to earn above-average returns; (3) I take average

[4] We use a temperature parameter of 1 and GPT-4o mini in our main analysis.

financial risks expecting average returns; and (4) I am not willing to take any financial risks. The four options are numbered from one to four, with higher numbers indicating greater levels of risk aversion. This self-assessment question is widely used as a proxy for risk aversion, particularly in financial decision-making. While it doesn't capture the entire spectrum of risk tolerance, it is consistent over time and correlates well with other measures of risk aversion obtained from hypothetical gambles and portfolio choices (Grable and Lytton 2001; Hanna and Lindamood 2004).

Our second dataset is the HILDA Survey in Australia, a nationally representative survey conducted annually since 2001. The HILDA Survey collects data on each person's risk preferences, cognitive abilities, personality traits, and identifies the household financial head. Risk aversion in the HILDA Survey is measured similarly to the SCF. For cognitive ability, respondents rate their math skills on a 0-10 scale relative to the average Australian adult. The standardized responses form a scale of math skills, which we use as a proxy for cognitive ability.

For Germany, our analysis utilizes the GSOEP survey, a longitudinal and nationally representative survey conducted annually since 1984. Our sample includes households from the years 2002, 2007, and 2012, which provide detailed information on demographics and risk tolerance. In the GSOEP, respondents report their willingness to take financial risks using an 11-point scale, where zero represents complete unwillingness to take risks and 10 signifies complete willingness. The specific question asked is: "How do you see yourself: Are you generally a person who is fully prepared to take risks, or do you try to avoid taking risks? You can use the values in between to make your estimate:

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10."

3.2. Methodology: Prompt Engineering

We selected an unambiguous prompt to prevent potential execution issues. Crafting the final prompt for the main results was an iterative process. To eliminate AI biases inherent to different countries, we evaluated ChatGPT 4's performance using three surveys from the USA, Australia, and Germany. Recognizing that our empirical setting uses English, which is theoretically easier for ChatGPT, we accounted for documented Western, especially American, biases in LLMs (Cao et al. 2023). To examine any country-specific biases, we included two additional surveys conducted in different countries. We also specified the precise year the human

respondent participated. Each survey's predefined questions provided various demographic information about respondents. We described the persona along five demographic dimensions: gender, age, income, education attainment, and employment status, across the USA (SCF survey), Australia (HILDA survey), and Germany (GSOEP survey).

To rigorously test which persona aspects most affect the accuracy of synthetic data, we collected 10 synthetic responses per human respondent using the following prompt. Additionally, we asked the AI to provide explanations for its responses and to rate its confidence in them. Our main results are based on the baseline version of the prompt, where ChatGPT adopts a persona defined by specific demographic values corresponding to a real survey respondent in each query as follows:

SCF Prompt:

"It is [YEAR]. You are a [AGE] year-old, [MARST], [GENDER] with [EDUC] years of education, making $ [INCOME] per year, living in the United States. You are [WORK] with [HEALTH] health status. Your occupation classification is [OCCAT2], and the number of correct answers on three financial literacy questions is [FINLIT]."

"Provide responses from this person's perspective.

Which of the statements on this page comes closest to the amount of financial risk that you are willing to take when you save or make investments?

1 if Take substantial financial risks expecting to earn substantial returns,

2 if Take above average financial risks expecting to earn above average returns,

3 if Take average financial risks expecting to earn average returns, and

4 if Not willing to take any financial risks.

Respond with these numbers, explanation, and confidence, separated by '|'."

HILDA Prompt:

"It is [YEAR]. You are a [AGE] year-old, [MARST], [GENDER] with [EDUCATION] years of education, making $ [INCOME] per year, living in Australia. You are [WORK] with [HEALTH] health status, standardized scores of self-reported math skills relative to the average or typical Australian adult is [COGNITIVE], and the number of correct answers on five financial literacy questions is [FINLIT]."

"Provide responses from this person's perspective.

Which of the following statements comes closest to describing the amount of financial risk that you are willing to take with your spare cash (i.e., cash used for savings or investment)?

1 if I take substantial financial risks expecting to earn substantial returns,

2 if I take above average financial risks expecting to earn above average returns,

3 if I take average financial risks expecting to earn average returns, and

4 if I am not willing to take any financial risks.

Respond with these numbers, explanation, and confidence, separated by '|'."

GSOEP Prompt:

"It is [YEAR]. You are a [AGE] year-old, married [GENDER] with [EDUCATION] years of education, making € [INCOME] per year, living in Germany. You are [WORK] with [N_CHILDREN] children, and your total wealth is [TOTAL_WEALTH]."

"Provide responses from this person's perspective.

How would you rate your risk tolerance at investments?

On a scale from 0 to 10, where 0 is not at all willing to take risks and 10 is very willing to take risks, what number would you be on the scale?

Respond with these numbers, explanation, and confidence, separated by '|'."

We generated responses from 10 synthetic respondents for each of the 5,000 human respondents randomly selected across the three surveys, resulting in a final dataset of 149,691 responses after data cleaning. This accounts for 99.79% of the expected data points (10×5,000×3 = 150,000). For each response, we recorded the numeric risk aversion level, the explanation provided by the LLM for the chosen score, and a measure of the model's reported confidence.

In our main analyses, we utilize all responses from the 10 synthetic respondents per human respondent, except when calculating uncertainty, where we rely only on the first synthetic response. To simplify comparison, we categorize the 10-level risk aversion measure into a 4-level measure based on the method by Kim, Hanna, and Ying (2021) as follows: 10 = 4; 9, 8, 7 = 3; 6, 5, 4 = 2; and 3, 2, 1 = 1.

3.3. Summary Statistics

Table 1 provides summary statistics for the variables used in the regressions. To ensure consistency, we normalize the healthy and financial literacy variables. As shown in Table 1, the

average risk aversion is 3.33 for humans and 2.98 for the LLM responses (including 10 synthetic responses), indicating that GPT generally tends to be more risk-seeking than humans.

Figure 1 illustrates the distribution of risk aversion levels among the survey respondents and their 10 synthetic responses generated by GPT. While the peak risk aversion (RA) for humans is at level 4, GPT's peak is at level 3, indicating that a significant portion of GPT responses reflect relatively lower risk aversion.

4. Results

We assess the quality of synthetic data through three approaches. First, we conduct univariate tests, such as analyzing correlations between Human RA and GPT RA, and comparing means and variances. While ChatGPT performs reasonably well in capturing overall variations, it often displays biases and shows considerable sensitivity to the personas represented in the prompts. Second, we examine whether ChatGPT can recover marginal associations between covariates, finding that regression results from human and synthetic samples often align, though sometimes with systematic differences. Third, we investigate its reasoning, noting that ChatGPT tends to be overly confident in its approximations of real human survey responses. Lastly, we validate our main results using different LLM models.

Despite differences in capturing the correct average of human data, we find that the overall average synthetic responses are close to population averages. Both human and LLM RA can be explained similarly by certain covariates in regressions. Thus, GPT-generated RA can be valuable for exploring preliminary results before conducting tests with real data. For the types of associational questions that interest economists, synthetic survey data performs well in terms of sign and statistical significance in regressions.

4.1. Univariate Tests

To assess whether the synthetic data's correlational structure matches the survey benchmarks, we examine the correlation between ChatGPT and human responses. Figure 2 presents a scatter plot where the scatter graph shows the average risk aversion (RA) from ChatGPT in each group of actual human RA levels. The gray line represents the linear polynomial fit based on the raw observations. The graph indicates a positive correlation between average human and ChatGPT RA. Interestingly, for lower levels of human RA (1 or 2 points), ChatGPT exhibits higher

RA than humans. Conversely, for higher levels of human RA (3 or 4 points), ChatGPT displays lower RA than humans.

Additionally, we compute the correlations between human and ChatGPT-generated RA. Table 2 presents Pearson correlations among the sample variables. GPT's RA is positively correlated with actual human RA, age, and the indicator for female gender. It is negatively correlated with income, education, health status, and the SCF survey indicator, suggesting that the GPT persona for US participants tends to be more risk-seeking. In contrast, the GPT persona for German participants tends to be more risk-averse, while correlations with other demographic variables are relatively low. These correlation coefficients are highly statistically significant at the 5% level.

Panel B of Table 2 shows the mean RA values for GPT and human respondents, t-statistics for their differences, and median values alongside the p-value of the Wilcoxon test for distribution differences. The mean GPT RA is slightly lower than the human RA (2.98 vs. 3.33), indicating that GPT is generally more risk-seeking, while the median values are identical.

To enhance understanding, we present graphical evidence of how RA levels differ between human and ChatGPT responses across the three surveys. Figure 3 plots the sample means and standard deviations for various RA levels, estimated using either actual survey respondents or the average of synthetic respondents from ChatGPT. Ideally, the distributions would be identical since the synthetic data were matched to survey respondents on selected covariates.

While the average ChatGPT responses do not exactly match the average survey responses, every synthetic mean falls within one standard deviation of the actual human average. Generally, ChatGPT's RA levels are lower than those of humans, indicating more risk-seeking behavior. The distribution of synthetic responses closely mirrors the variation in human responses, with similar standard deviations (0.72 vs. 0.77). The variation in ChatGPT responses reflects statistical uncertainty due to sampling from the language model, making their similarity to human data notable.

For the SCF survey in the U.S., the human RA score is slightly higher than ChatGPT's level. In the HILDA survey in Australia, human and ChatGPT levels are nearly identical, with overlapping confidence intervals suggesting strong agreement. However, in the GSOEP survey in Germany, the human and ChatGPT levels differ the most. This visual comparison highlights the alignment and discrepancies between human and ChatGPT RA levels across different surveys.

Overlapping confidence intervals in all three surveys suggest that ChatGPT's RA levels are consistent with human responses, indicating potential reliability in modeling human-like RA behavior.

Despite synthetic data performing well overall, issues arise when examining subgroups. We calculate average RA variations across gender and age groups, split into two main columns for "female" and "male," and further divided by age groups: 18-24, 25-34, 35-54, 55-74, and 75+. Each group is analyzed using the three surveys. Figure 4 compares RA levels across different age and gender groups between human and ChatGPT responses. ChatGPT tends to be more risk-seeking across all age and gender groups, especially for young and male individuals. Both human and ChatGPT mean RAs are close for elderly groups. Notably, ChatGPT exaggerates RA for females and older individuals. For example, in the 75+ age range in the GSOEP survey, scores for both genders are almost identical. In terms of gender differences, the standard deviation for male participants is relatively larger than for females. When examining the data by education category, males with a bachelor's degree or higher exhibit over 0.25 points more RA than females in the same group. On the 1 to 4-point RA scale (1 = least risk-averse, 4 = most risk-averse), a difference of more than 0.25 points indicates a significant difference in RA between groups. In general, Figure 4 highlights that synthetic responses suggest the AI agent is more risk-averse for specific demographic groups such as females and the elderly than it actually is. These groups are generally found to be more risk-averse, as documented by prior research. On average, ChatGPT is more risk-seeking than actual humans for male, young, educated, high-income, and employed personas.

4.2. Multivariate Tests

We further examine the differences in regression results when using true and synthetic risk aversion as the dependent variable. We estimate linear regression models as follows:

$$Y_{i(c),t} = \alpha_c + \alpha_t + \beta \times X_{i,t} + \varepsilon_{c,t} \quad (1)$$

where i indexes respondents at a survey year t in a survey c, X_i is a vector of the respondent characteristics used in our persona prompt (age, gender, education, income, employment status, marital status, health status, and financial literacy), and $Y_{i(c),t}$ represents risk aversion for human or GPT. We are most interested in the vector of β coefficients measuring how the partial correlation between each covariate in X_i and the risk aversion differs between the synthetic responses and the human benchmark.

Table 3 presents the results. The coefficients for all demographic characteristics are significant and in the same direction for GPT. Specifically, females, older individuals, those with lower income, less education, unemployed, unmarried, lower financial literacy, and unhealthy individuals are more risk-averse. However, the test for equality of coefficients generally fails, indicating a significant difference between the coefficients for humans and GPT, except for health status. [5]

Overall, we conclude that ChatGPT's risk aversion is highly correlated with human risk aversion, accurately reflecting that females, older individuals, those with lower income, less education, unemployed, unmarried, lower financial literacy, and unhealthy individuals are more risk-averse. However, the economic magnitudes of these correlations may differ.

5. Validation

Our main results relied on ChatGPT's numeric responses to prompts about expressing risk aversion on a scale from 1 (or 0) to 4 (or 10). Numbers can be challenging for a language model to generate consistently because they exist in a similar embedding space, making random selection more likely. This occurs because tokens in close proximity in the embedding space have similar posterior probabilities. Despite this, we observed consistent patterns across covariates, which is reassuring. Nonetheless, validating these values is essential.

To validate these values, we prompted ChatGPT to provide short explanations for its chosen risk aversion levels. We found surprising consistency in these explanations, allowing us to document correlations between keywords in the explanations and the risk aversion levels. Keywords are defined based on demographics reported in Table A1. Figure 5 plots the most commonly occurring keywords (those used 100 or more times) and their associated risk aversion levels, with point sizes indicating the number of instances of each explanation-RA level pair. This plot provides reassuring validation in two ways. First, keywords associated with high risk aversion levels (e.g., "age," "wealth," "income," "ill," and "financial literacy") are sensible.

Second, we used a BERT transformer to embed these explanations in semantic space, producing numeric representations of each explanation. If the AI uses risk aversion levels

[5] The insignificance of income in the human data is attributed to multicollinearity with education, as income and education are highly correlated. In un-tabulated results where we exclude education, we find that lower income is significantly associated with higher risk aversion.

accurately, we would expect that the differences between explanations associated with a level 1 and a level 2 are similar to those between a level 3 and a level 4, and both smaller than the differences between a level 1 and a level 4. We measure the difference between risk aversion levels using an absolute difference metric and the differences between explanations using a cosine similarity metric of each explanation's embeddings. We calculate the average embedding for each risk aversion level-by-survey-by-gender, then collapse the data to the average cosine distance for each risk aversion difference-by-gender.

The second graph in Figure 5 visualizes the results with risk aversion differences on the x-axis and cosine distances between embeddings on the y-axis, with each facet representing a survey-by-gender. The plots trend upward in HILDA and SCF, indicating that more significant differences in numeric risk aversion responses are accompanied by more significant differences in explanations in semantic space. However, the plot trends linearly in GSOEP, suggesting that ChatGPT may struggle to provide valid explanations for chosen risk aversion levels for German respondents.[6]

Third, we used ChatGPT to help annotate a random SCF sample of 200 pairs of explanations, inspired by Beckmann et al. (2024). In each comparison, we asked ChatGPT to indicate which explanation justified a certain level of risk aversion. The prompt is as follows:

"Please read the following justifications of an individual's choice, which were given in response to the following question:

Which of the following statements comes closest to describing the amount of financial risk that you are willing to take when you save or make investments?

1 if Take substantial financial risks expecting to earn substantial returns,

2 if Take above average financial risks expecting to earn above average returns, and

3 if Take average financial risks expecting to earn average returns, and

4 if Not willing to take any financial risks.

What are high-level categories to choose 1, 2, 3, or 4?

Make sure that each justification can be assigned to one of the categories.

Respond with these numbers and high-level categories, separated by '|'. Here is the justification: [Explanation] "

[6] One of possible differences between HILDA/SCF and GSOEP is language, where the former uses English while the latter uses German.

We present the frequency distribution of Human RA, GPT RA (used in our baseline results), and GPT RA2 (GPT-generated RA obtained in the annotation step) in Table 4. We begin by comparing the predicted RA levels from both GPT-generated datasets to the actual human risk aversion. The results show weak performance overall, with only 35% (690 out of 1,988) of GPT RA matching human RA. However, there is a high degree of agreement between the two GPT-generated RA levels, with 91% of them being identical. This strong correlation suggests a robust association between the GPT-generated RA levels and their reasoning.

To further validate the risk aversion levels produced by ChatGPT, we employ Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to generate topics from each explanation. We estimate 28 topics (*k*) using both unigrams and bi-grams on the explanations, producing a matrix of document-topic weights $\theta_{d,k}$ describing how likely it is that explanation *d* is about topic *k*, as well as a matrix of word-topic weights $\varphi_{w,k}$ that describe how likely it is that word *w* is associated with topic *k*.[7] We use these values to calculate the weighted average risk aversion score per topic by weighting each synthetic response by the topic distribution θ across explanations. Formally:

$$Avg\ Risk\ aversion_k = \sum_d \theta_{d,k} * Risk\ aversion_d \quad (2)$$

As illustrated in the third graph of Figure 5, there is a clear association between the topics (indicated on the y-axis by the top 5 most associated words, determined by the ϕ word-topic distribution) and the risk aversion levels most commonly found in the explanations linked to those topics. For instance, higher risk aversion levels are associated with topics related to “income”, “wealth” and include words such as "avoid" and "jeopardize." In contrast, lower risk aversion levels are described using terms like "confidence," "earn_above," "background," and "opportunities."

6. Confidence, uncertainty and distribution of GPT

ChatGPT is a generative language model that optimizes to predict the most likely next word based on previous context. The temperature hyperparameter in the model controls the level of determinism in this prediction process. Higher temperatures allow the model to choose not just the highest probability word but also lower probability words, adding variability to its responses.

[7] We first estimate 30 topics and drop 2 topics which were just type errors. We then remove any typos in the top 5 terms.

This implies that the results from our prompts are characterized by their own probability distribution. While our main findings indicated that the model's estimates were overly precise compared to human responses, the exact nature of this probability distribution remains unclear.

6.1. Confidence

Although ChatGPT doesn't provide direct access to the posterior probabilities of its outputs, we experimented by asking the AI to indicate its confidence in the risk aversion levels it generated. Our prompts yielded responses that were mostly on a scale from 0% (not confident) to 100% (very confident), though the precise wording varied. In Figure 6, we observed few instances at the extremes of this scale, with most responses indicating levels like "fairly confident" or "feel confident." The second graph of Figure 6 displays the distribution of these confidence levels, with the majority indicating confidence levels above 70%.

We analyzed these self-reported confidence levels across two dimensions: survey type and gender. The third graph of Figure 6 illustrates that the average confidence for responses from the GSOEP survey was lower than those from the SCF or HILDA surveys, with HILDA responses showing the highest confidence. This variation is sensible as the GSOEP survey provides less detailed persona descriptions, leading to lower confidence. This effect was more pronounced for female personas, suggesting that less detailed persona sketches result in decreased confidence in risk aversion predictions.

When examining risk aversion itself, we observed that neutral risk aversion levels were associated with lower confidence, whereas extreme RA levels (either high or low) corresponded with higher confidence, forming a U-shaped pattern in the AI's confidence levels as shown in the fourth graph of Figure 6.

These analyses rely on the LLM's self-reported confidence to gauge the variability of the synthetic data. The last graph of Figure 6 shows a negative relationship between self-reported confidence and empirical uncertainty (measured as the standard deviation of RA levels), suggesting that higher confidence correlates with lower variation in the data. This indicates that self-reported confidence is a reasonable proxy for empirical uncertainty.

6.2. Uncertainty

To understand the uncertainty in AI-generated data, we measured the standard deviation of RA levels. ChatGPT produced less variation compared to real humans, as documented in Table 1. A possible explanation is that less frequently occurring groups are harder for the algorithm to predict due to their smaller representation in the training data. With 30 samples per human-by-survey, we analyzed the standard deviation at the level of human-by-survey profiles, averaging these deviations by the number of humans in each profile, combining five demographics: gender, age, income, education, and employment.

The first graph of Figure 7 shows little evidence that ChatGPT is more confident for commonly occurring groups, contrary to expectations. The second graph shows that male respondents had slightly larger standard deviations, except in the SCF survey, where the average standard deviation for the most common group was slightly less than 0.65, compared to over 0.7 for the least common group. [8] These synthetic data deviations are more precise than those observed in real human data.

We also explored whether these patterns reflect only sample sizes or demographic stereotypes (e.g., older females being more risk-averse). Variation in human RA levels is driven by genuine differences in risk aversion among specific groups. We found less evidence of these relationships in the synthetic data, suggesting that demographic factors, rather than sparsely populated profiles, pose a greater challenge. The third graph of Figure 7 plots standard deviations for ChatGPT profiles with at least 30 human respondents, showing that LLM estimates are generally more precise than human data, regardless of sample size.

Finally, we present descriptive plots of average per-persona standard deviations, broken down by survey (fourth graph) and covariates (fifth graph). Despite some variability across surveys, synthetic data consistently exhibited less noise than human data.

6.3. Distribution

A crucial point of interest is the distributions. With 10 synthetic respondents per human respondent, we aggregate all covariates and plot the empirical distributions by profile and survey. Figure 8 displays the distribution for synthetic risk aversion (RA) levels. Horizontal bars represent the range of LLM RA values, vertical marks indicate the median, and colors differentiate the

[8] The randomly picked distribution is similar to that of all human SCF surveys from 1992 to 2022, showing female to male ratio is around 30%.

gender of the persona. The plot reveals highly consistent estimates of variance, with most RA ranges spanning between 2 and 4, and rarely extending to the lowest value of 1.

For instance, older females without a college education tend to have higher median RA levels. The second graph illustrates similar patterns across different surveys. Notably, while the overall support for RA levels is generally higher for females, the GSOEP survey shows the most distribution among female respondents.

7. Research Design Choices

7.1. Temperature

In the context of ChatGPT, the term "temperature" refers to how deterministic the choice of the next token in its output is, based on the prompt and preceding tokens. The model selects each subsequent token from a probability distribution. A temperature setting of zero ensures that the AI will always choose the highest probability token. As the temperature increases, the selection process becomes less deterministic, introducing variability often referred to as "creativity."

Our primary results were generated with ChatGPT set to its most "creative" mode, with the temperature hyperparameter set to 1. The default temperature settings for ChatGPT and Llama are 0.7 (ranging from 0 to 1), while Gemini defaults to 1.0 (ranging from 0 to 2). Technically, the temperature can exceed 1. In theory, higher temperatures should produce noisier results that more closely resemble the actual randomness in human survey responses. However, even at this high setting, our findings indicate that ChatGPT's estimates were more precise than those of human respondents, with ChatGPT's estimates having a lower average standard deviation (0.72 compared to 0.77 for humans).

This standard deviation combines two sources of variation. The first is the variation from averaging across different demographic groups, such as age, gender, educational attainment, employment status, and income. The second source is the inherent randomness of the data generation process. For humans, this randomness reflects individual quirks not captured by covariates. In ChatGPT, it is an intrinsic characteristic of the model, influenced by the temperature setting.

To explore how standard deviation might increase with reduced creativity, we calculated the standard deviation for each profile in the LLM data at temperature settings ranging from 0 to 1, using a consistent prompt. In Figure 9, we observed that adding more detailed demographics

reduced the average standard deviation of RA levels across LLMs. Notably, the most detailed demographic groups at a temperature setting of 1 showed the highest standard deviation. Table 5 confirms that a temperature of 1 yielded the highest proportion of correctly generated LLM risk aversion levels. Therefore, we adopted a temperature setting of 1 for our baseline results.

7.2. Other AI Agents

Our primary analysis utilizes OpenAI's GPT-4o-mini model. In this section, we also examine the performance of OpenAI's GPT-4o, Google's Gemini 2.0 Flash (experimental), and Meta's Llama 3.3 70B (versatile) large language models. [9] Table 5 shows that GPT, in contrast to Gemini and Llama, consistently delivers superior results across all temperature settings. The best result is achieved with GPT at a temperature of 1, with an accuracy of 49%. For our main results, we chose the cost-efficient GPT-4o-mini model, which demonstrates 41% accuracy at a temperature of 1.0.

Figure 10 summarizes the updated performance of GPT in terms of basic correlations with the SCF data, relative to the GPT-4o-mini version, revealing a few improvements. Specifically, GPT-4o has a lower intercept of 1.3 (compared to 1.9) and a higher slope of 0.46 (compared to 0.34), where the human data represents a zero intercept and a slope of 1.0. However, the overall performance relative to GPT-4o-mini is not economically significant. Considering the substantial cost difference between the two models, the slightly superior performance does not alter our main conclusion. Despite its demonstrated superior performance across various tasks, improvements in ChatGPT's underlying language model have not significantly enhanced its ability to generate synthetic data for the application we examine.

8. Discussion and Conclusion

The rapid advancement of artificial intelligence has introduced new possibilities for data generation and analysis. This study investigates the efficacy of data generated by ChatGPT in replicating human survey responses, with a specific focus on risk aversion (RA) metrics. Through a series of comprehensive tests, we evaluate the quality of AI-generated data by comparing it with human data across several dimensions. Our analysis reveals that while ChatGPT can replicate

[9] These models are the most recent and powerful models at the time of writing this draft from the companies.

overall variations and capture marginal associations between covariates, it often introduces biases and exhibits sensitivity to the specific personas represented in the prompts.

To validate the reasoning processes of ChatGPT, we examine the consistency of its explanations for RA level choices, demonstrating a strong correlation between keywords and specific RA levels. We also explore the confidence levels of AI-generated RA data and investigate how these relate to the observed variability. Our findings indicate a negative relationship between self-reported confidence and empirical uncertainty, suggesting that ChatGPT's confidence is a reasonable proxy for the variability of its outputs.

Additionally, we address the uncertainty of AI-generated data by examining the standard deviation of RA levels and comparing these with human data. Our results show that ChatGPT generates less variation compared to real human data, possibly due to the algorithm's difficulty in predicting less frequently occurring groups. We also explore the distributions of synthetic RA levels, highlighting the reliability of AI-generated data for certain demographic groups.

Finally, we discuss the impact of research design choices, such as the model's temperature setting and the selection of AI agents, on the performance of ChatGPT. While AI-generated data shows promise for preliminary research and hypothesis testing, several limitations must be considered, including the model's tendency to overfit certain demographic profiles and inherent biases in the training data.

References

Beckmann, Lars, Heiner Beckmeyer, Ilias Filippou, Stefan Menze, and Guofu Zhou, 2024, Unusual financial communication-evidence from chatgpt, earnings calls, and the stock market .

Bisbee, J., Clinton, J.D., Dorff, C., Kenkel, B. and Larson, J.M., 2024. Synthetic replacements for human survey data? The perils of large language models. Political Analysis, 32(4), pp.401-416.

Brand, J., Israeli, A. and Ngwe, D., 2023. Using GPT for market research. Harvard Business School Marketing Unit Working Paper, (23-062).

Bybee, J Leland, 2023, The ghost in the machine: Generating beliefs with large language models, arXiv preprint arXiv:2305.02823 .

Campbell, J.Y., 2006. Household finance. The journal of finance, 61(4), pp.1553-1604.

Cao, Y., L. Zhou, S. Lee, L. Cabello, M. Chen, and D. Hershcovich. 2023. "Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study." In Proceedings of the FirstWorkshop on Cross-Cultural Considerations in NLP (C3NLP), 53–67. Dubrovnik: Association for Computational Linguistics. https://aclanthology.org/2023.c3nlp-1.7.

Chang, Anne, Xi Dong, and Changyun Zhou, 2024, Does democratized information access demand democratized ai? a tale of the rich, the poor, and the ai (chatgpt) in earnings calls.

Chen, Y., Liu, T. X., & others. (2023). The emergence of economic rationality of GPT. Proceedings of the National Academy of Sciences.

Crosetto, P., & Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. Experimental Economics, 19, 613-641.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.

Dillion, D., Tandon, N., Gu, Y. and Gray, K., 2023. Can AI language models replace human participants?. Trends in Cognitive Sciences, 27(7), pp.597-600.

Eisfeldt, A.L. and Schubert, G., 2024. AI and Finance (No. w33076). National Bureau of Economic Research.

Eisfeldt, Andrea L, Gregor Schubert, Miao Ben Zhang, and Bledi Taska, 2023, The labor impact of generative ai on firm values, Available at SSRN 4436627 .

Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. Management Science, 69(4), 1935-1950.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier, 2024, Chatgpt and perception biases in investments: An experimental study, Available at SSRN 4787249 .

Giannetti, M. and Wang, T.Y., 2016. Corporate scandals and household stock market participation. The Journal of Finance, 71(6), pp.2591-2636.

Grable, J.E. and Lytton, R.H., 2001. Assessing the concurrent validity of the SCF risk tolerance question. Journal of Financial Counseling and Planning, 12(2), p.43.

Gu, R., Peng, C. and Zhang, W., 2024. The Gender Gap in Household Bargaining Power: A Revealed-Preference Approach. Review of Financial Studies.

Hanna, S.D. and Lindamood, S., 2004. An improved measure of risk aversion. Journal of Financial Counseling and Planning, 15(2), pp.27-45.

Hewitt, Luke, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer, 2024, Predicting results of social science experiments using large language models, Unpublished working paper.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. American economic review, 92(5), 1644-1655.

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

Jha, Manish, Jialin Qian, Michael Weber, and Baozhong Yang, 2024, Chatgpt and corporate policies, Technical report, National Bureau of Economic Research.

Jia, J., Yuan, Z., Pan, J., McNamara, P., & Chen, D. (2024). Decision-Making Behavior Evaluation Framework for LLMs under Uncertain Context. arXiv preprint arXiv:2406.05972.

Johnson, J.E. and Powell, P.L., 1994. Decision making, risk and gender: Are managers different?. British journal of management, 5(2), pp.123-138.

Kim, K.T., Hanna, S.D. and Ying, D., 2021. The Risk Tolerance Measure in the 2016 Survey of Consumer Finances: New, But Is It Improved?. Journal of Financial Counseling & Planning, 32(1).

Lo, A.W. and Ross, J., 2024. Can ChatGPT Plan Your Retirement?: Generative AI and Financial Advice. Working Paper.

Morin, R.A. and Suarez, A.F., 1983. Risk aversion revisited. The journal of finance, 38(4), pp.1201-1216.

Ranjan, R., Gupta, S. and Singh, S.N., 2024. A comprehensive survey of bias in llms: Current landscape and future directions. arXiv preprint arXiv:2409.16430.

Tranchero, Matteo, Cecil-Francis Brenninkmeijer, Arul Murugan, and Abhishek Nagaraj, 2024, Theorizing with large language models, Technical report, National Bureau of Economic Research.

Van Rooij, M., Lusardi, A. and Alessie, R., 2011. Financial literacy and stock market participation. Journal of Financial economics, 101(2), pp.449-472.

FIGURE 1: Distribution of ChatGPT Persona and Actual Humans' Risk Aversion

This figure presents a density plot of the risk aversion of human respondents and ChatGPT.

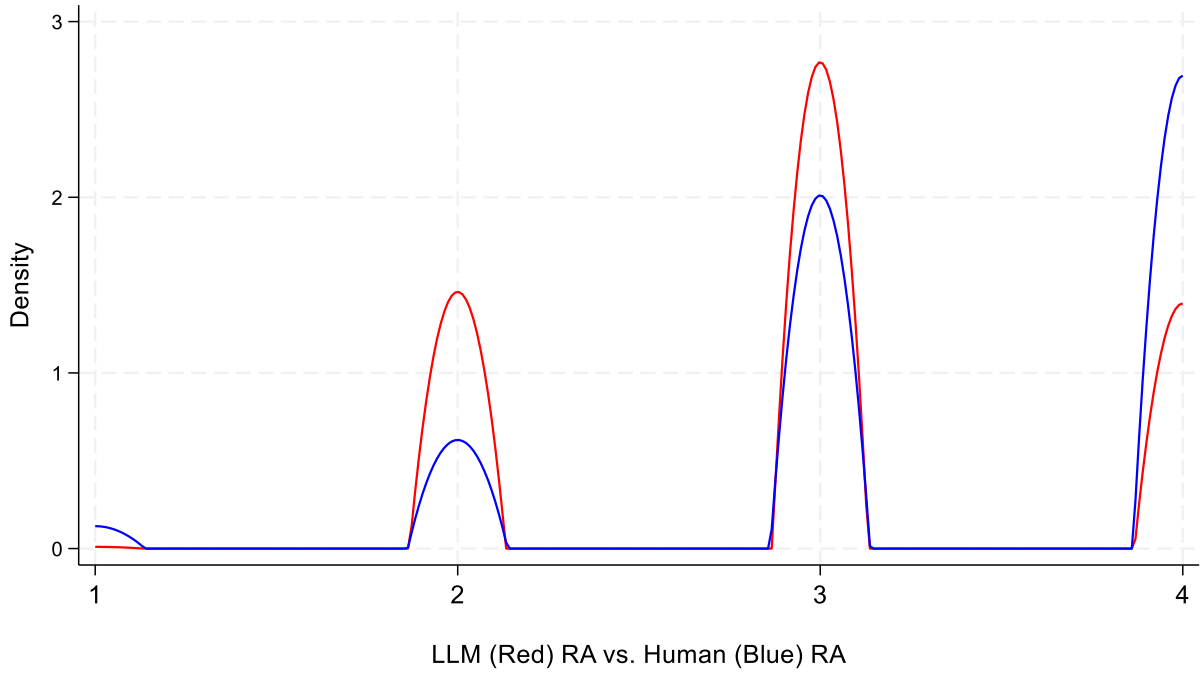


FIGURE 2: Relation between ChatGPT Persona and Actual Humans' Risk Aversion

This figure presents a scatter plot of the correlation between the risk aversion of human respondents (x-axis) and ChatGPT (y-axis). The scatter graph shows the average within the group. The line represents the linear polynomial fit using the underlying raw observations.

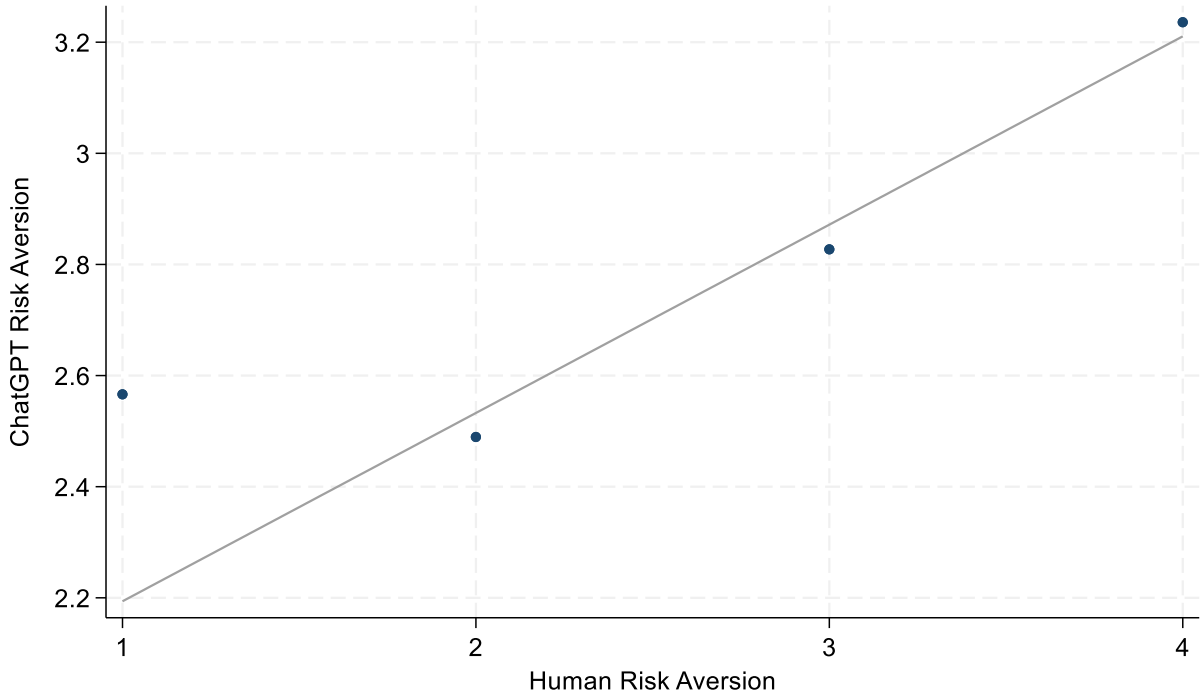


FIGURE 3: Comparison of Risk Aversion between ChatGPT Persona and Actual Humans

The figure shows a comparison of risk aversion scores between human respondents and ChatGPT across three different surveys: SCF, HILDA, and GSOEP. Average actual human respondents' responses from SCF, HILDA, and GSOEP surveys indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars.

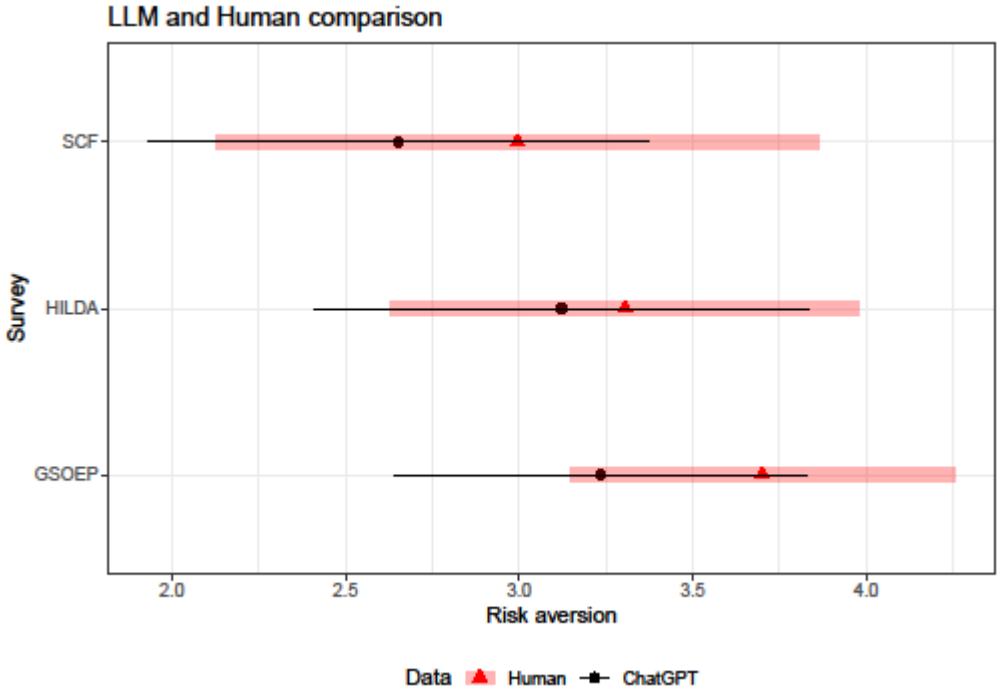
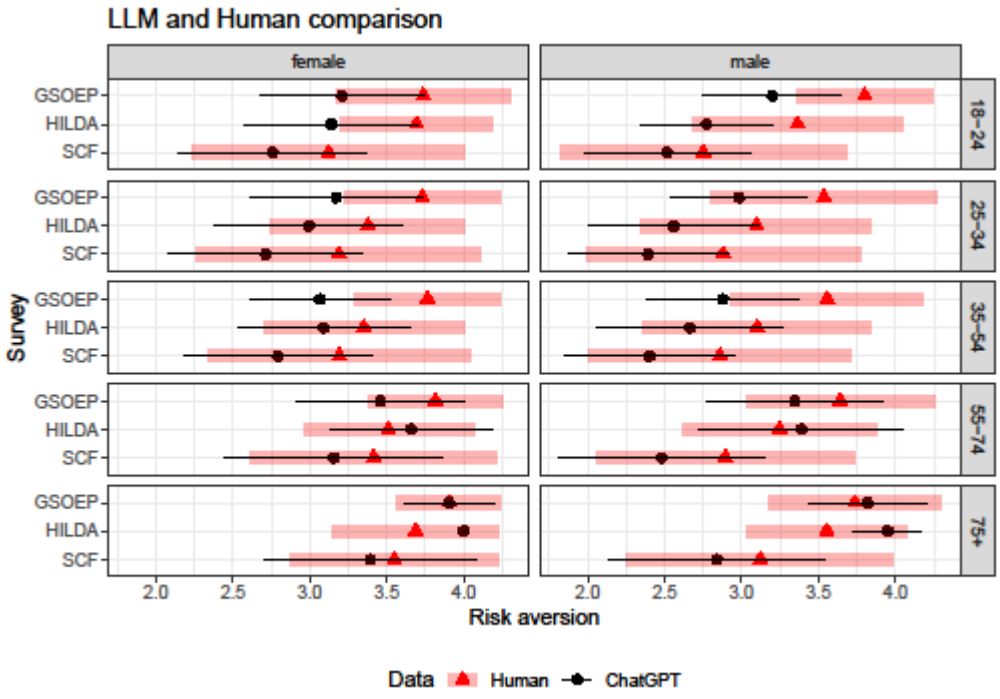
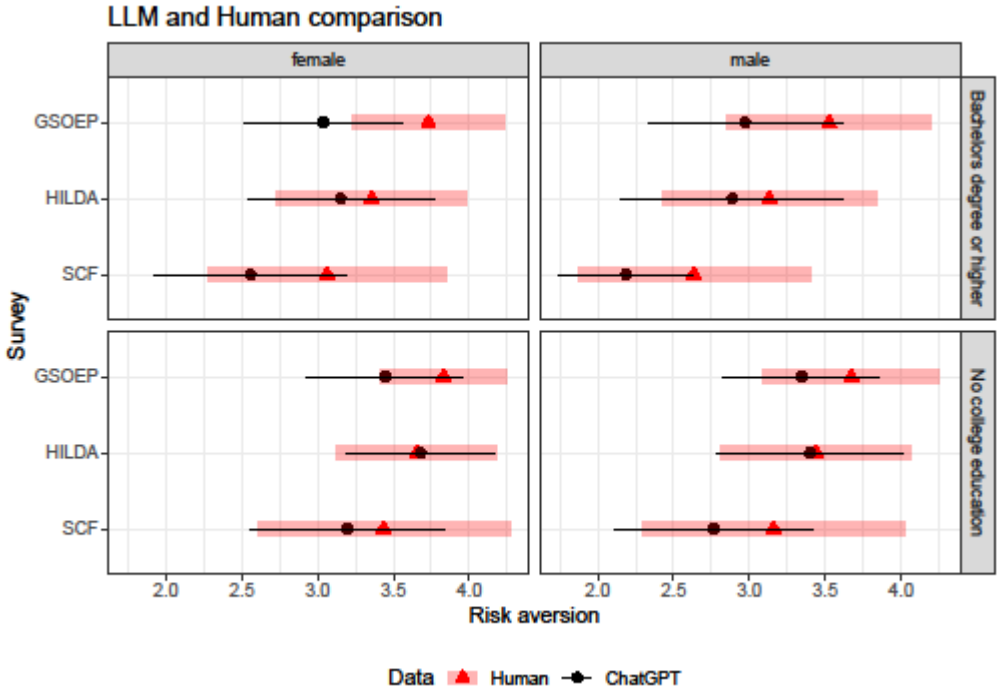
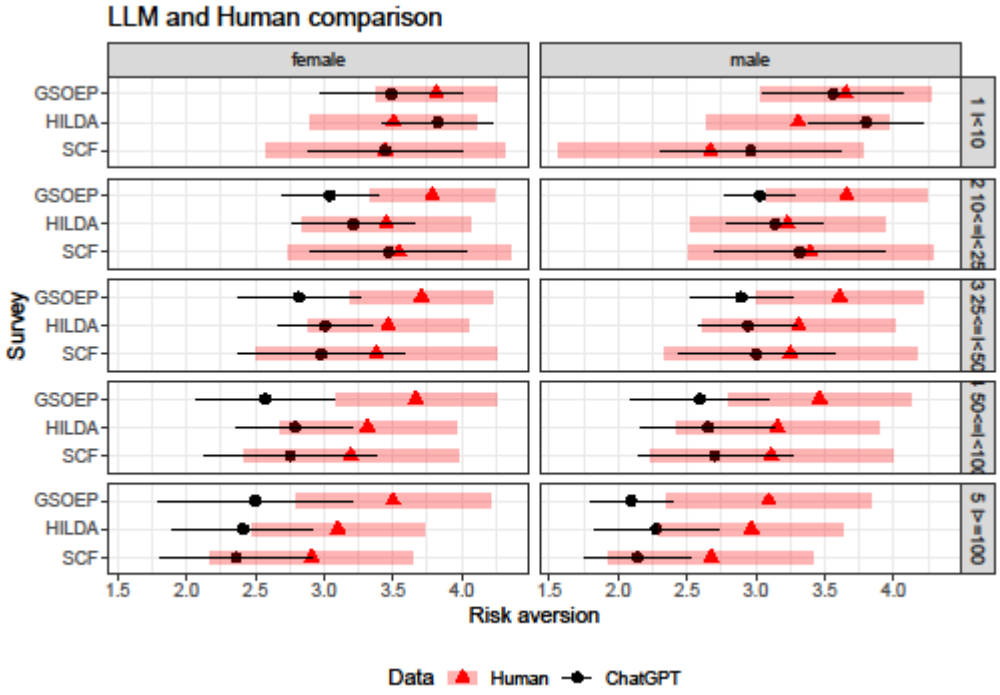


FIGURE 4: ChatGPT Persona vs. Actual Humans by Gender, Age, Education, Income, and Employment Status

The figure shows a comparison of risk aversion scores between human respondents and ChatGPT across three different surveys: SCF, HILDA, and GSOEP by gender, age/education attainment/income (in thousand)/employment status group. Average actual human respondents' responses from SCF, HILDA, and GSOEP surveys indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars.




LLM and Human comparison
female
male
Bachelors degree or higher
No college education
GSOEP
HILDA
SCF
Survey
2.0
2.5
3.0
3.5
4.0
Risk aversion
Data
Human
ChatGPT


LLM and Human comparison
female
male
1 k<10
2 10<=k<25
3 25<=k<50
4 50<=k<100
5 k>=100
GSOEP
HILDA
SCF
Survey
1.5
2.0
2.5
3.0
3.5
4.0
Risk aversion
Data
Human
ChatGPT

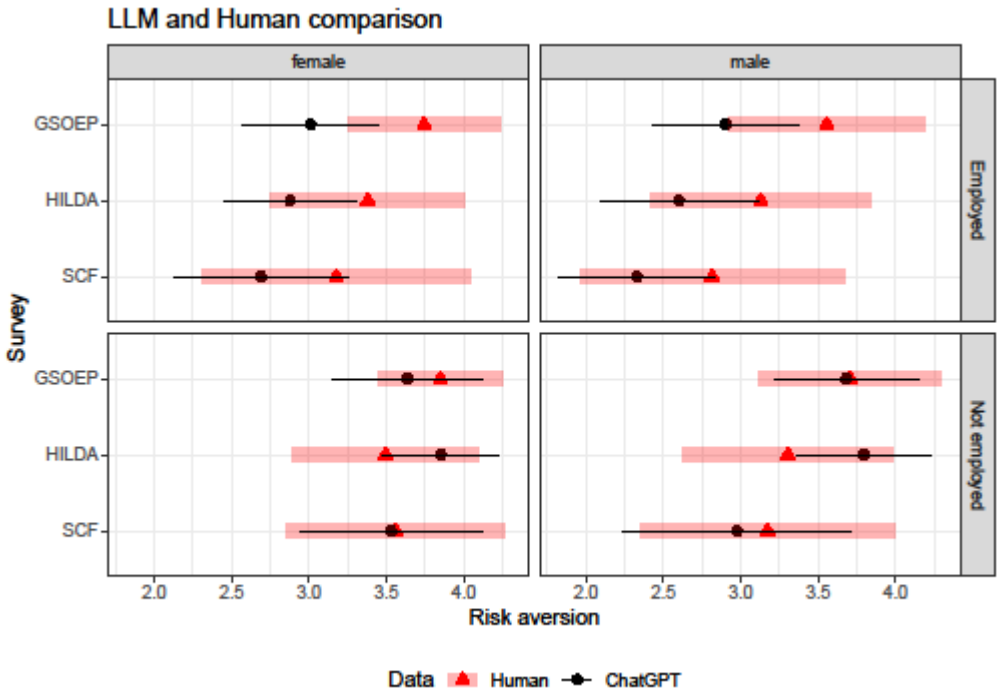
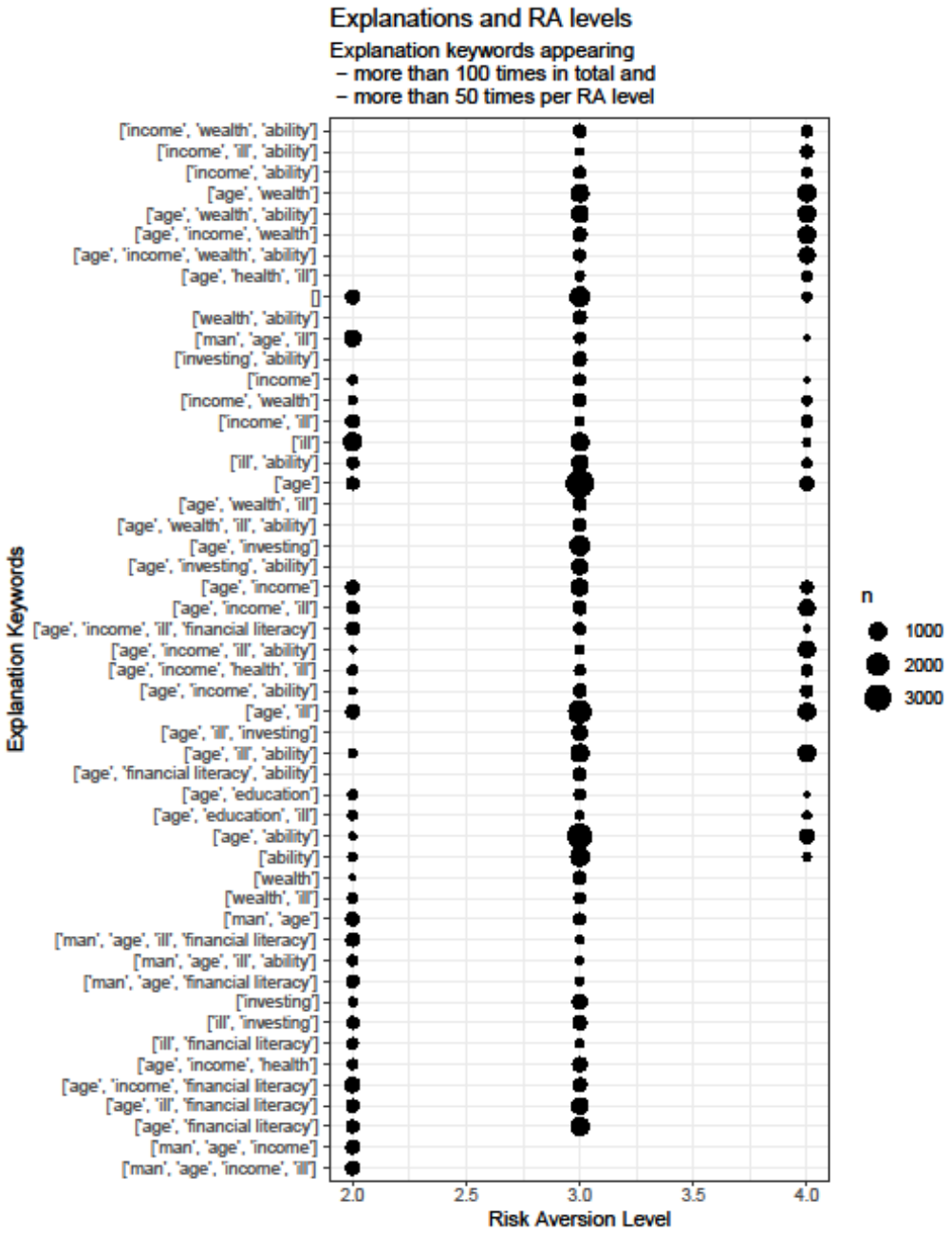
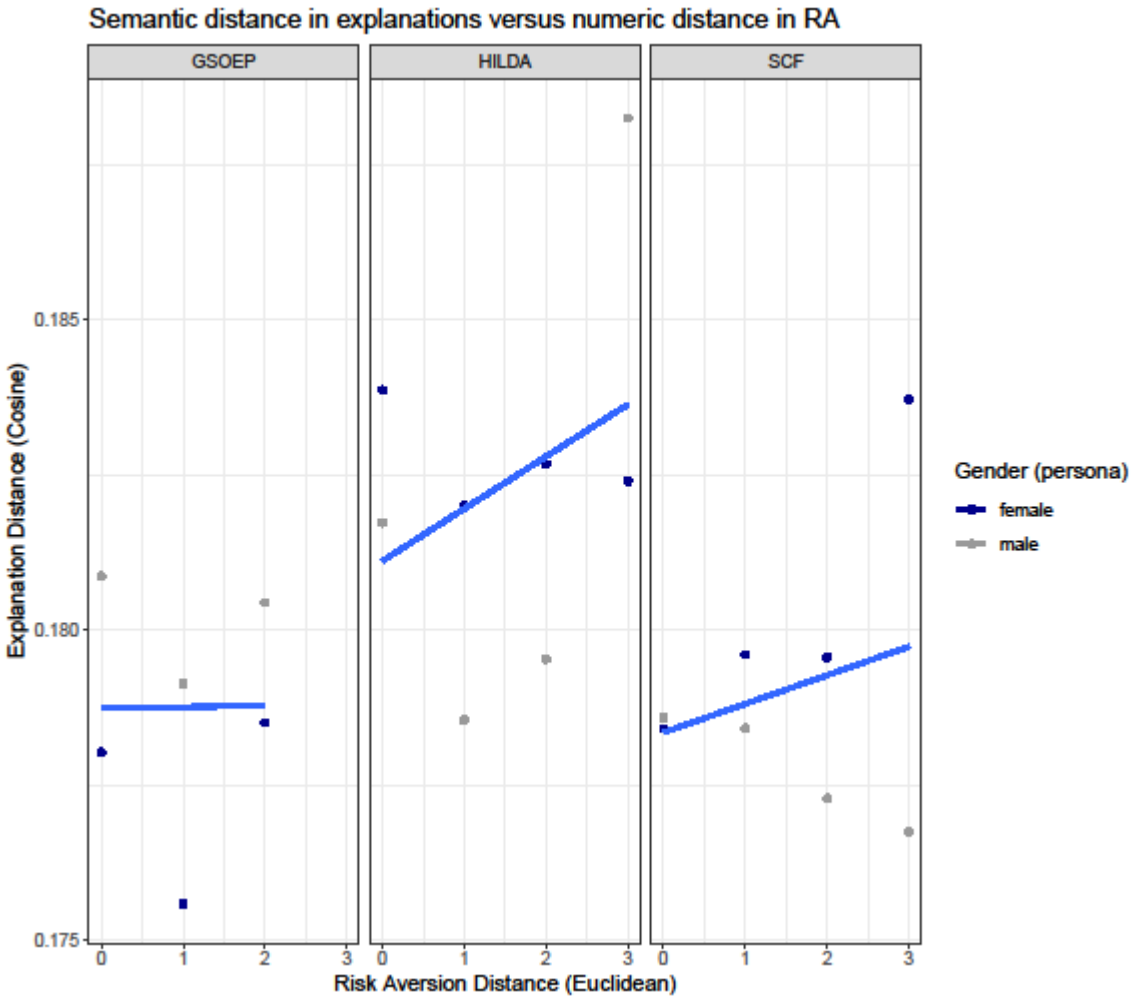

LLM and Human comparison
female
male
Employed
Not employed
GSOEP
HILDA
SCF
Survey
2.0
2.5
3.0
3.5
4.0
Risk aversion
Data
Human
ChatGPT

FIGURE 5: Explanation

The figures show various tests of GPT's self-reported explanation on why it chooses a certain level of risk aversion.




Semantic distance in explanations versus numeric distance in RA
GSOEP
HILDA
SCF
Explanation Distance (Cosine)
0.185
0.180
0.175
0
1
2
3
Risk Aversion Distance (Euclidean)
Gender (persona)
female
male

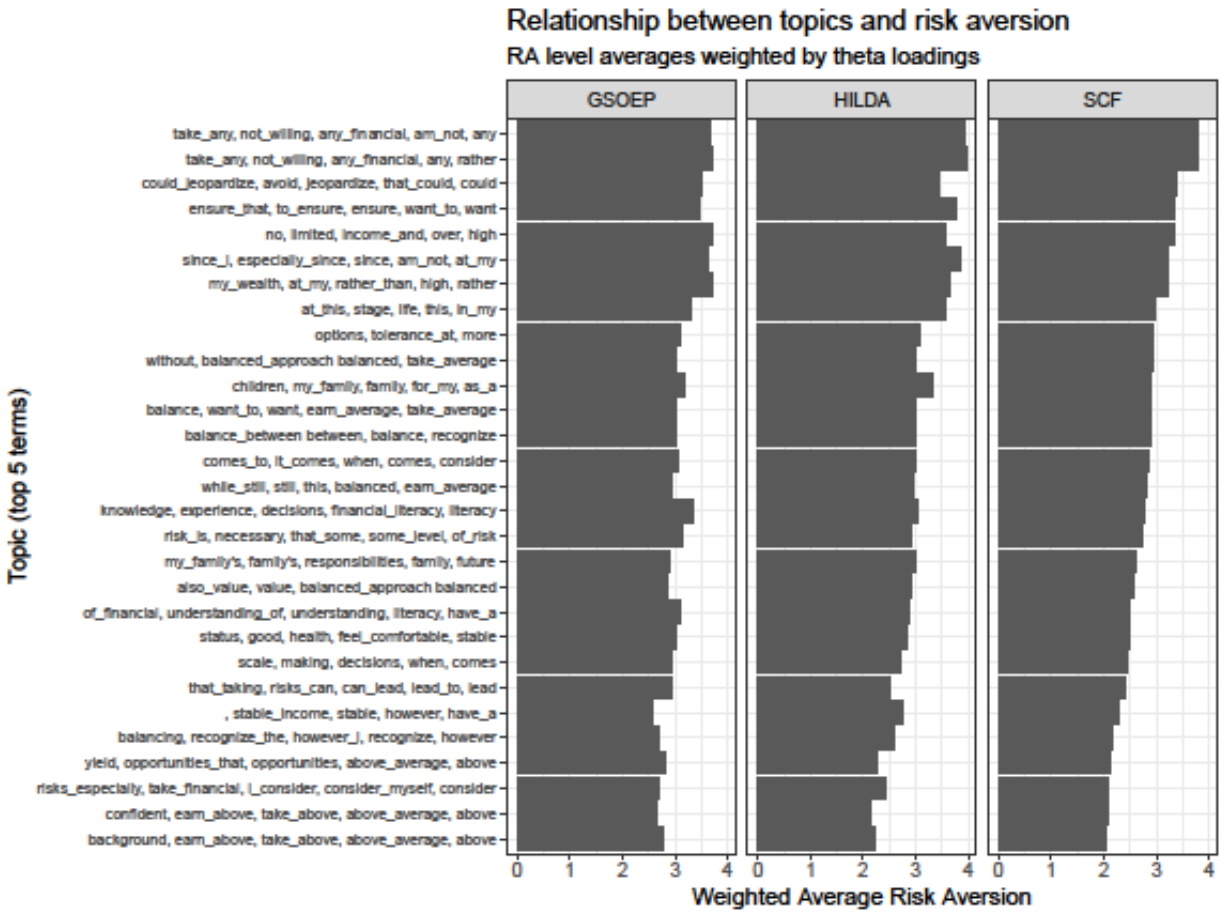
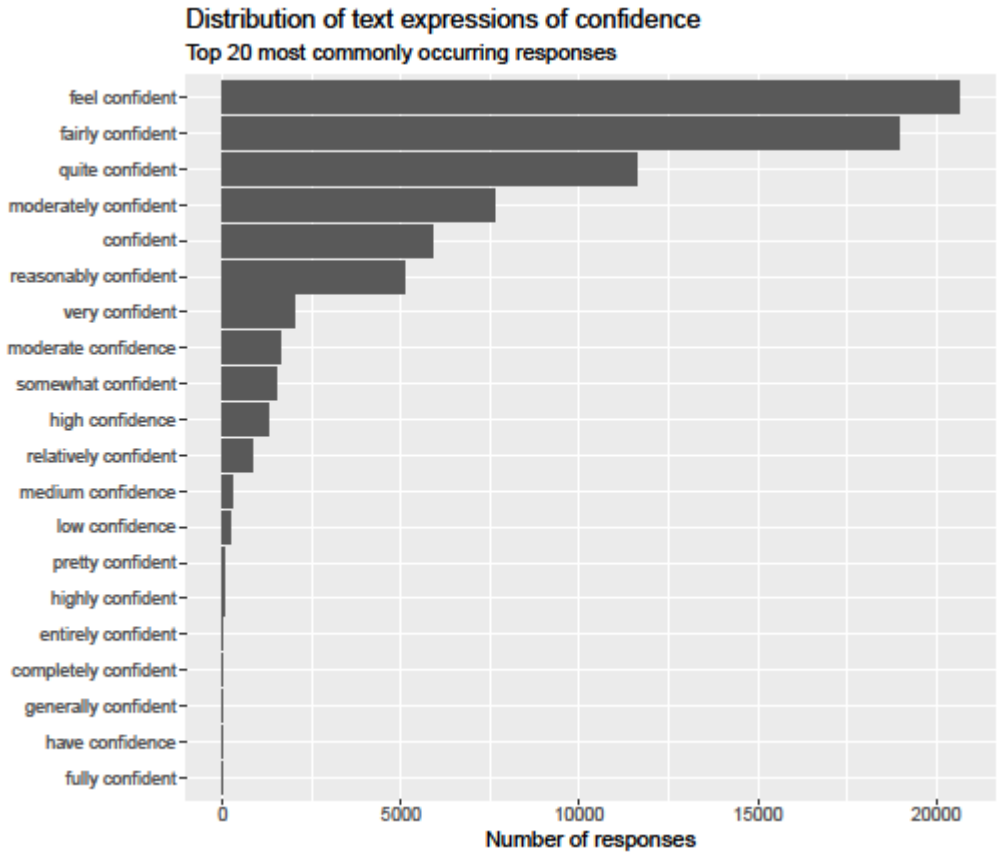
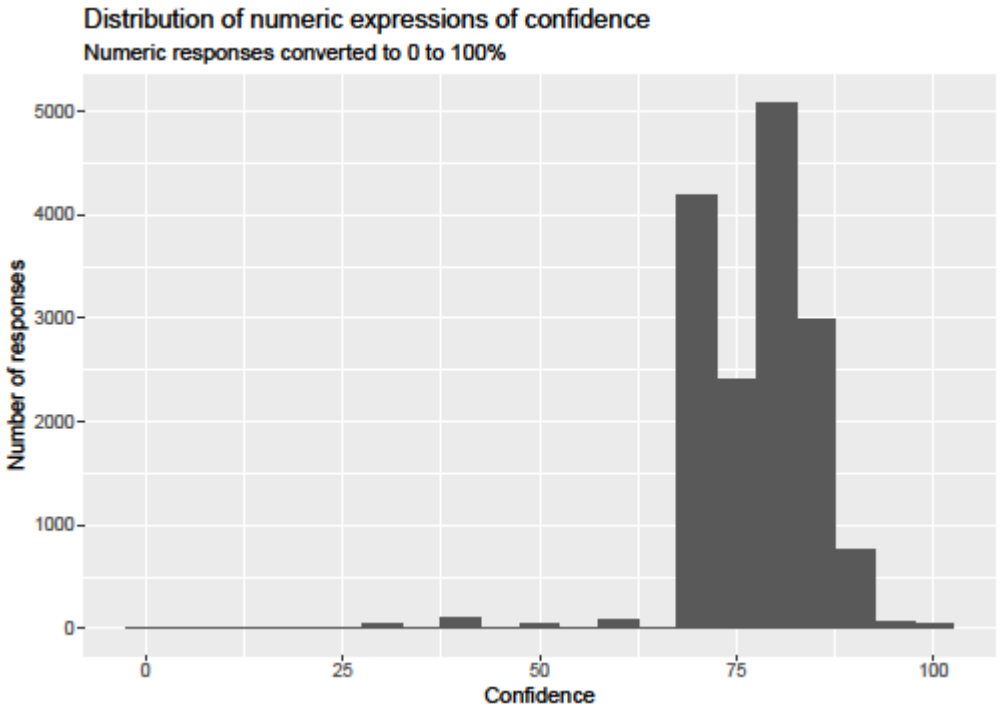
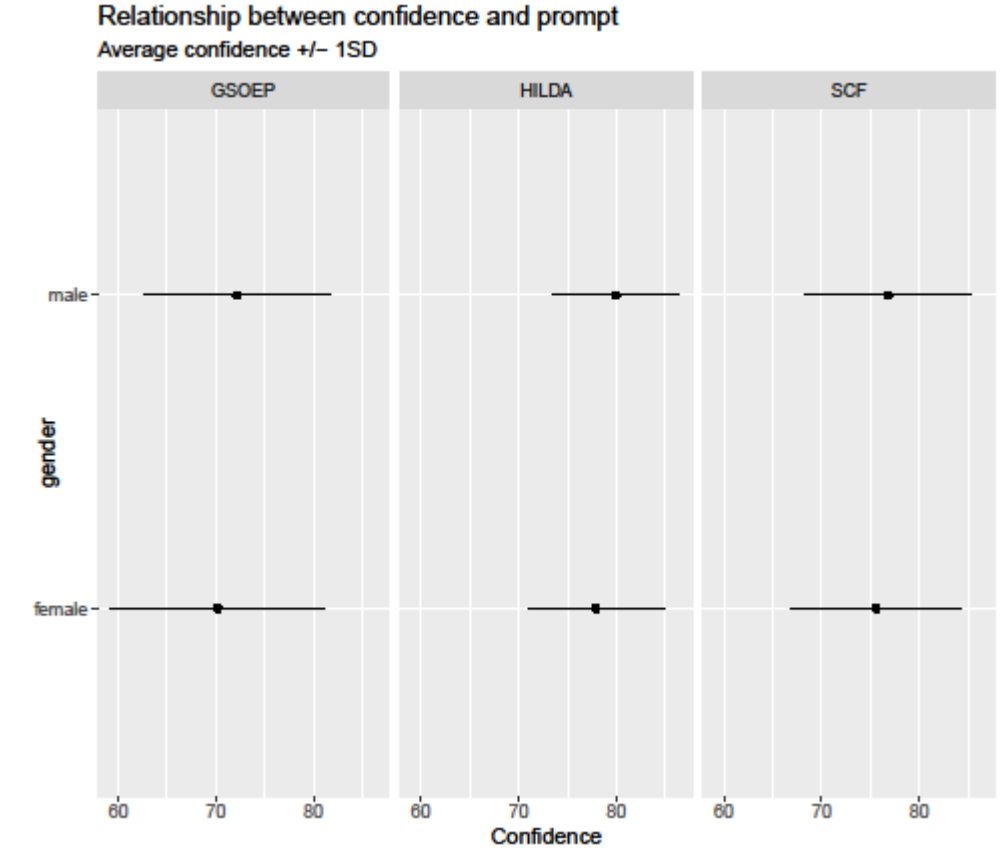

Relationship between topics and risk aversion
RA level averages weighted by theta loadings
GSOEP
HILDA
SCF
Topic (top 5 terms)
take_any, not_willing, any_financial, am_not, any
take_any, not_willing, any_financial, any, rather
could_jeopardize, avoid, jeopardize, that_could, could
ensure_that, to_ensure, ensure, want_to, want
no, limited, income_and, over, high
since_i, especially_since, since, am_not, at_my
my_wealth, at_my, rather_than, high, rather
at_this, stage, life, this, in_my
options, tolerance_at, more
without, balanced_approach balanced, take_average
children, my_family, family, for_my, as_a
balance, want_to, want, earn_average, take_average
balance_between between, balance, recognize
comes_to, it_comes, when, comes, consider
while_still, still, this, balanced, earn_average
knowledge, experience, decisions, financial_literacy, literacy
risk_is, necessary, that_some, some_level, of_risk
my_family's, family's, responsibilities, family, future
also_value, value, balanced_approach balanced
of_financial, understanding_of, understanding, literacy, have_a
status, good, health, feel_comfortable, stable
scale, making, decisions, when, comes
that_taking, risks_can, can_lead, lead_to, lead
, stable_income, stable, however, have_a
balancing, recognize_the, however_i, recognize, however
yield, opportunities_that, opportunities, above_average, above
risks_especially, take_financial, i_consider, consider_myself, consider
confident, earn_above, take_above, above_average, above
background, earn_above, take_above, above_average, above
0
1
2
3
4
Weighted Average Risk Aversion

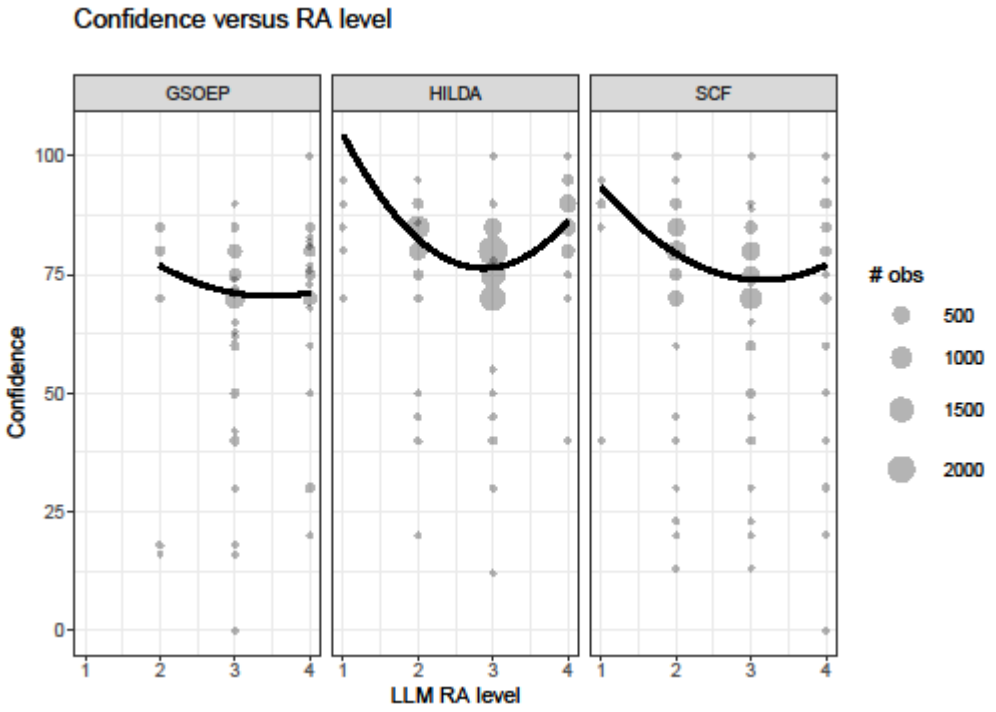
FIGURE 6: Confidence

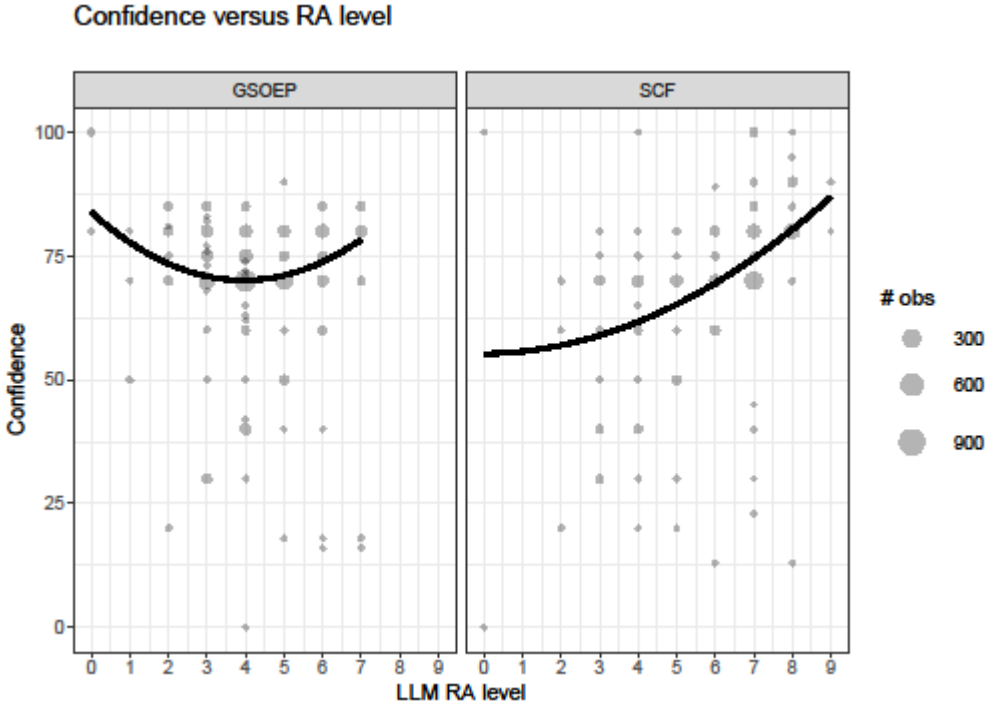
The figures show various tests of GPT's self-reported confidence.




Distribution of numeric expressions of confidence
Numeric responses converted to 0 to 100%
Number of responses
5000
4000
3000
2000
1000
0
0
25
50
75
100
Confidence


Relationship between confidence and prompt
Average confidence +/− 1SD
GSOEP
HILDA
SCF
gender
male
female
60
70
80
Confidence


Confidence versus RA level
GSOEP
HILDA
SCF
Confidence
LLM RA level
obs
500
1000
1500
2000


Confidence versus RA level
GSOEP
SCF
Confidence
LLM RA level
obs
300
600
900

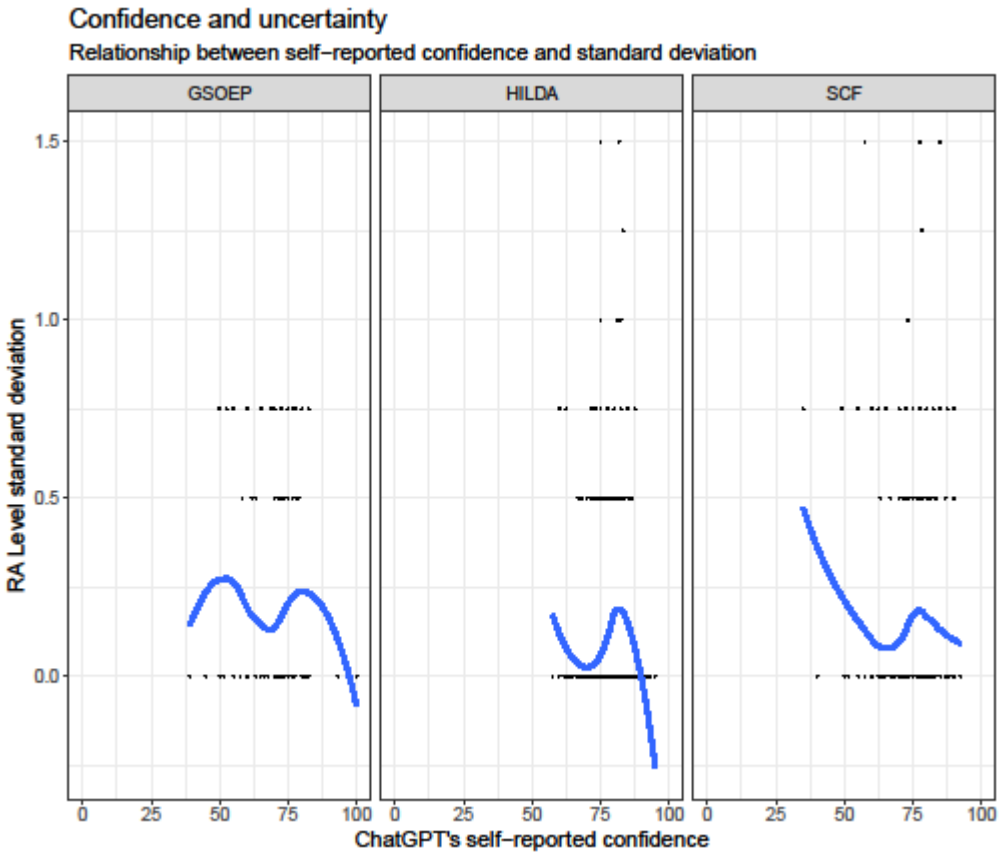
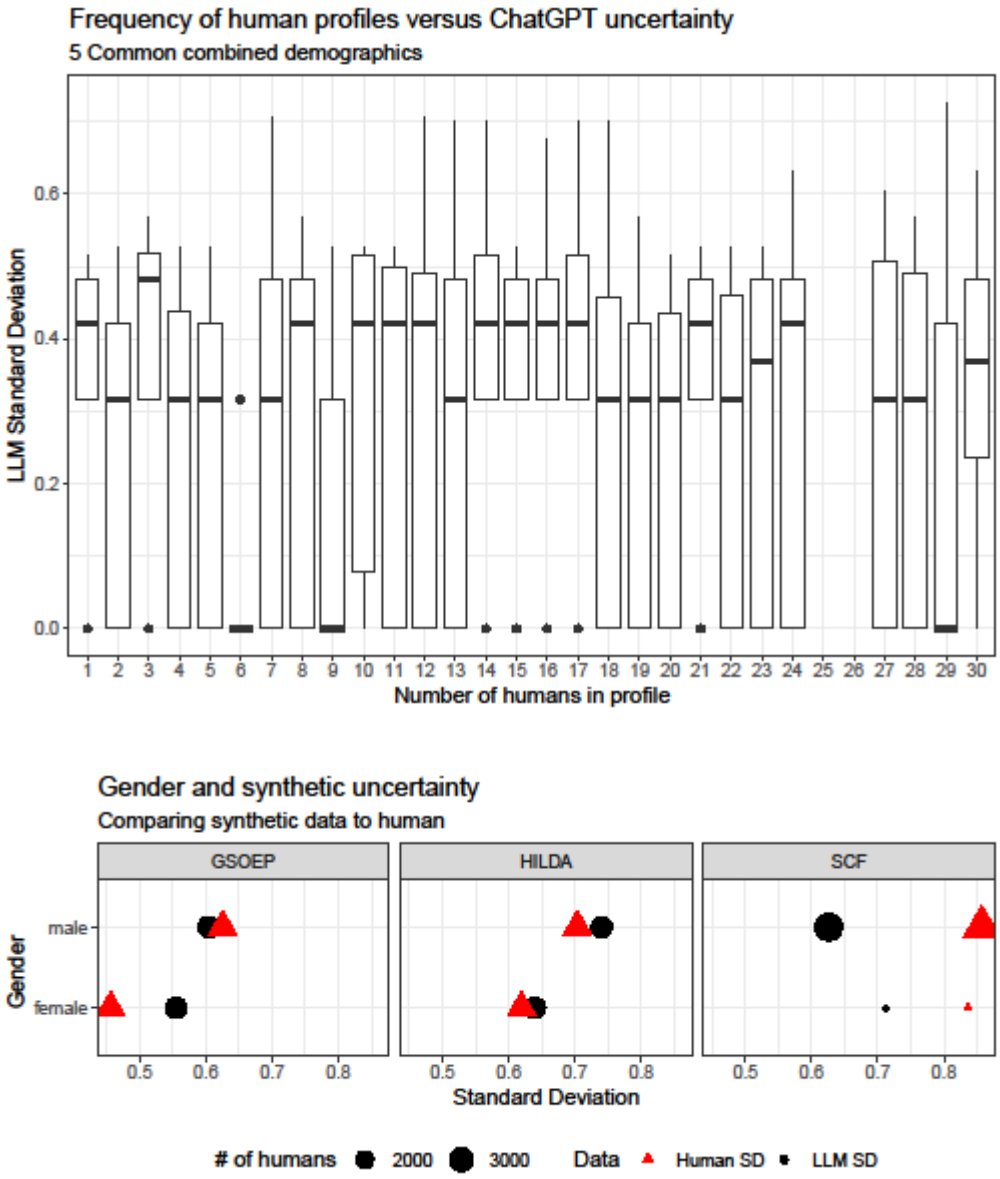
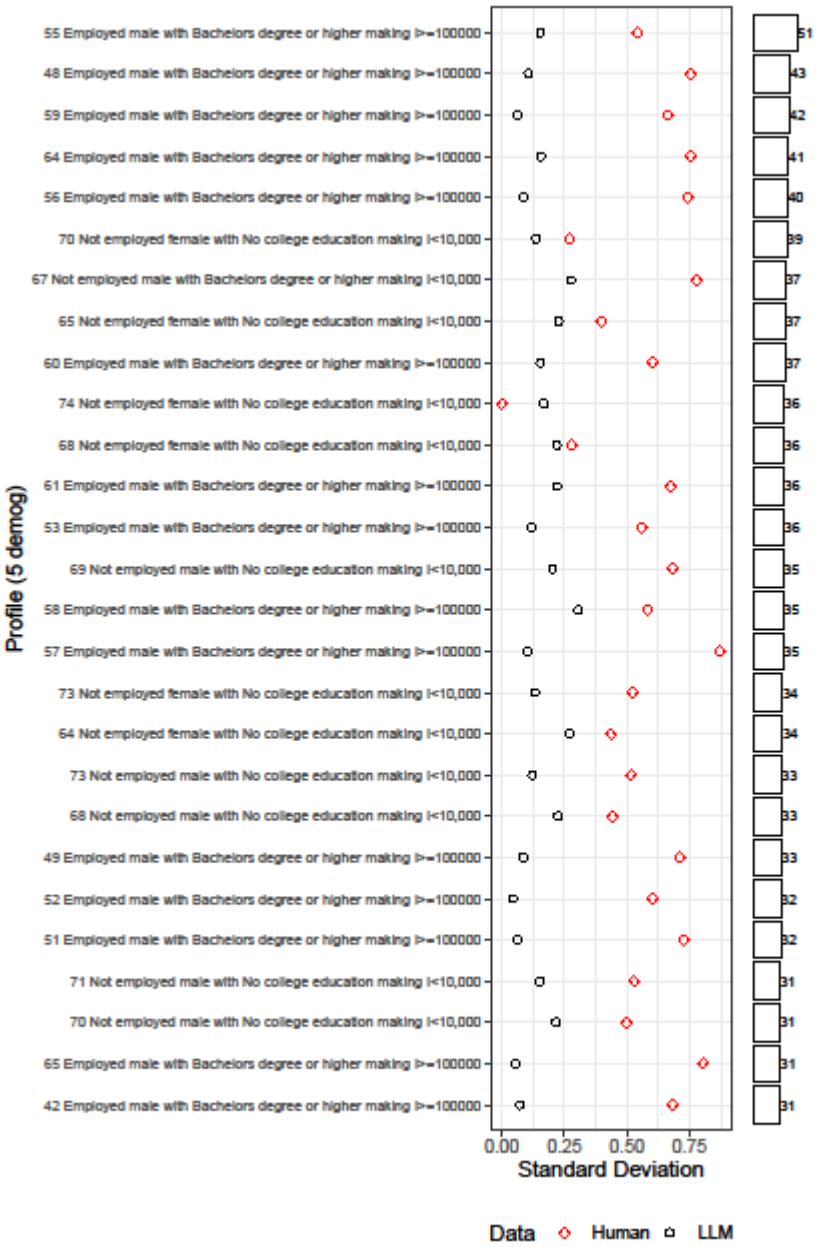
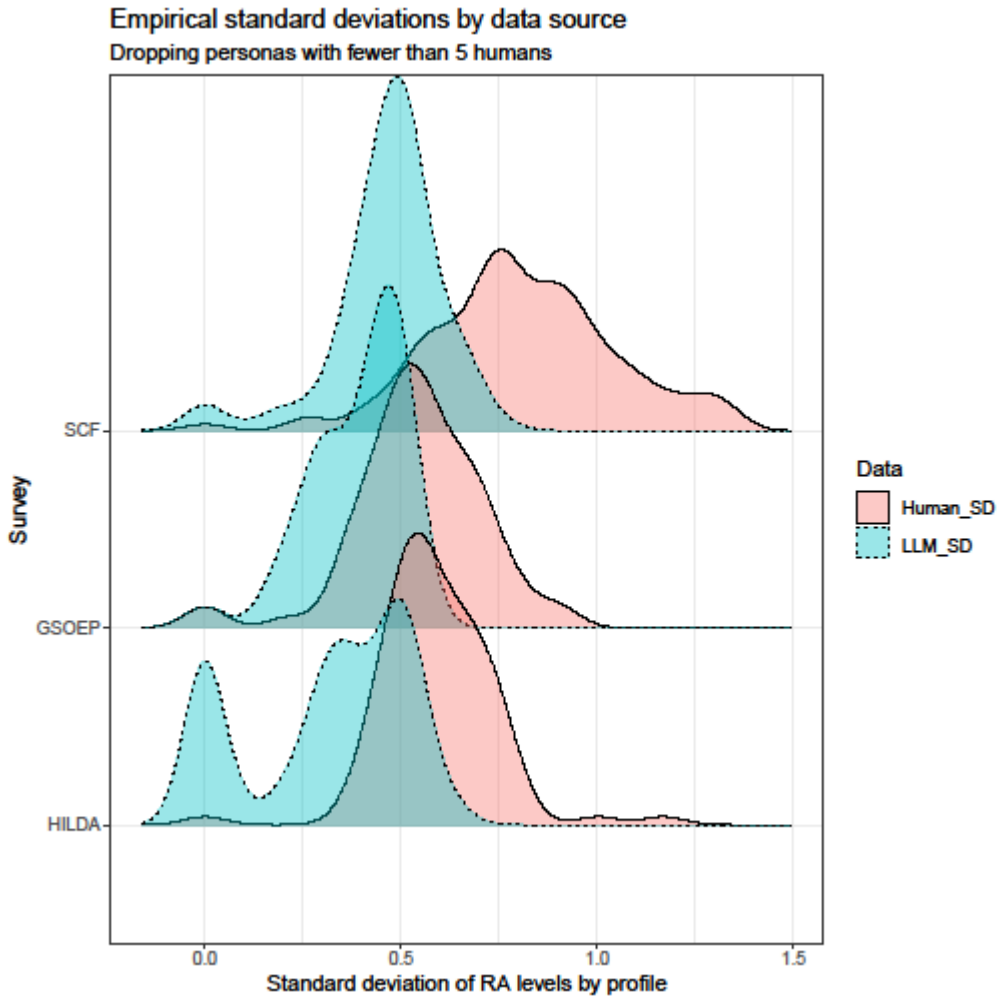

Confidence and uncertainty
Relationship between self−reported confidence and standard deviation
GSOEP
HILDA
SCF
RA Level standard deviation
1.5
1.0
0.5
0.0
0
25
50
75
100
ChatGPT's self−reported confidence

FIGURE 7: Uncertainty

The figures show various tests of standard deviation of GPT RA.




Profile (5 demog)
55 Employed male with Bachelors degree or higher making I>=100000
48 Employed male with Bachelors degree or higher making I>=100000
59 Employed male with Bachelors degree or higher making I>=100000
64 Employed male with Bachelors degree or higher making I>=100000
56 Employed male with Bachelors degree or higher making I>=100000
70 Not employed female with No college education making I<10,000
67 Not employed male with Bachelors degree or higher making I<10,000
65 Not employed female with No college education making I<10,000
60 Employed male with Bachelors degree or higher making I>=100000
74 Not employed female with No college education making I<10,000
68 Not employed female with No college education making I<10,000
61 Employed male with Bachelors degree or higher making I>=100000
53 Employed male with Bachelors degree or higher making I>=100000
69 Not employed male with No college education making I<10,000
58 Employed male with Bachelors degree or higher making I>=100000
57 Employed male with Bachelors degree or higher making I>=100000
73 Not employed female with No college education making I<10,000
64 Not employed female with No college education making I<10,000
73 Not employed male with No college education making I<10,000
68 Not employed male with No college education making I<10,000
49 Employed male with Bachelors degree or higher making I>=100000
52 Employed male with Bachelors degree or higher making I>=100000
51 Employed male with Bachelors degree or higher making I>=100000
71 Not employed male with No college education making I<10,000
70 Not employed male with No college education making I<10,000
65 Employed male with Bachelors degree or higher making I>=100000
42 Employed male with Bachelors degree or higher making I>=100000
51
43
42
41
40
39
37
37
37
36
36
36
36
35
35
35
34
34
33
33
33
32
32
31
31
31
31
0.00
0.25
0.50
0.75
Standard Deviation
Data
Human
LLM


Empirical standard deviations by data source
Dropping personas with fewer than 5 humans
Survey
SCF
GSOEP
HILDA
0.0
0.5
1.0
1.5
Standard deviation of RA levels by profile
Data
Human_SD
LLM_SD

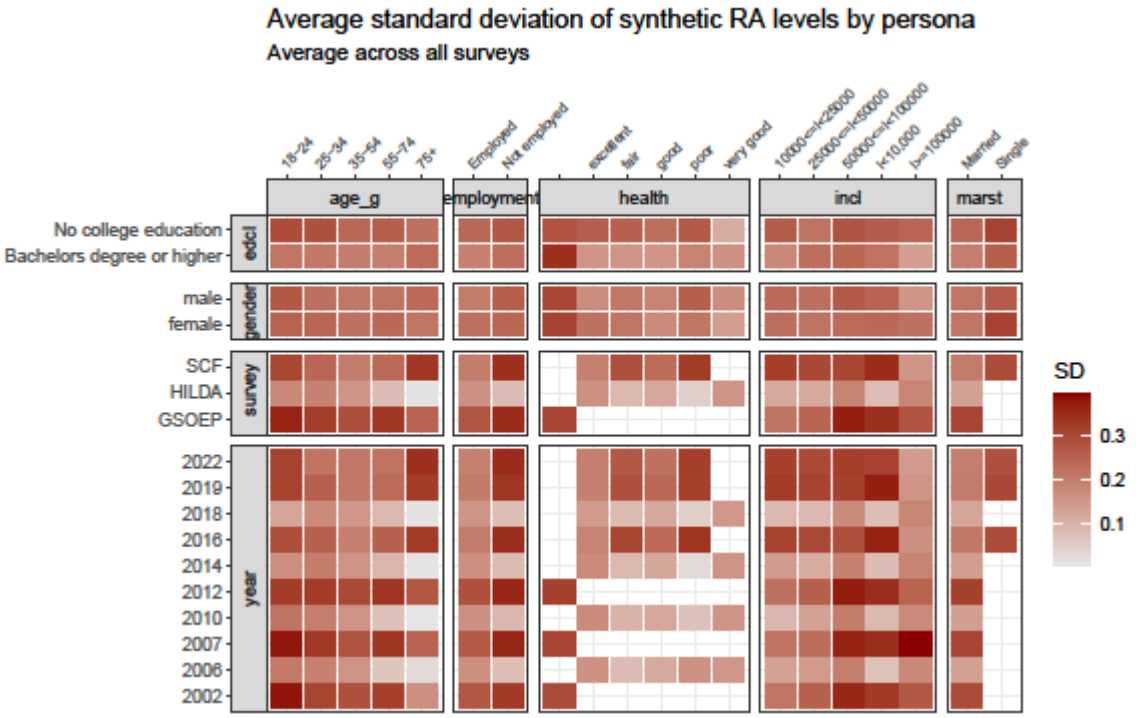
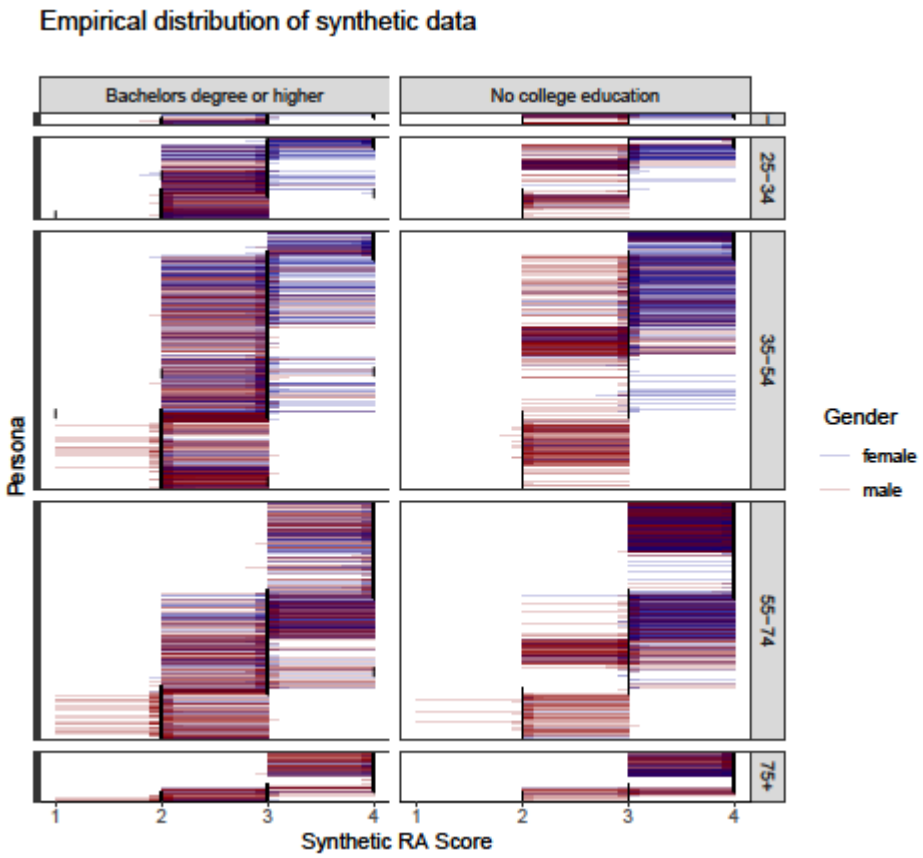

Average standard deviation of synthetic RA levels by persona
Average across all surveys
age_g
employment
health
incl
marst
18–24
25–34
35–44
45–54
55–74
75+
Employed
Not employed
excellent
fair
good
poor
very good
I<10,000
10,000<=I<25,000
25,000<=I<50,000
50,000<=I<100,000
I>=100,000
Married
Single
edcl
No college education
Bachelors degree or higher
gender
male
female
survey
SCF
HILDA
GSOEP
year
2022
2019
2018
2016
2014
2012
2010
2007
2006
2002
SD
0.3
0.2
0.1

FIGURE 8: Distribution

The figures show the distribution of GPT risk aversion level by gender by education by age group (or by survey).



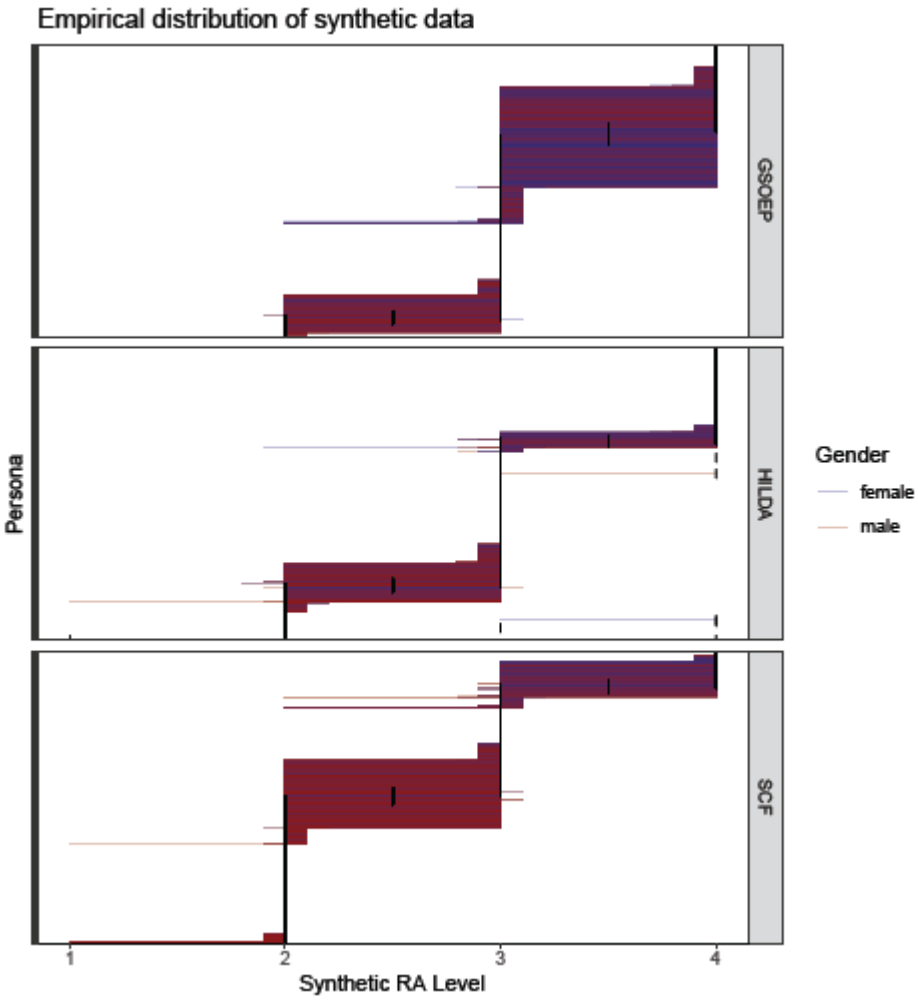

Empirical distribution of synthetic data
GSOEP
HILDA
SCF
Persona
Gender
female
male
1
2
3
4
Synthetic RA Level

FIGURE 9: Temperature

The figure shows the relationship between temperature hyperparameter (t=0, 0.7, or 1) and average standard deviation of LLM RA by LLM model (rows) and level of persona aggregation (columns).

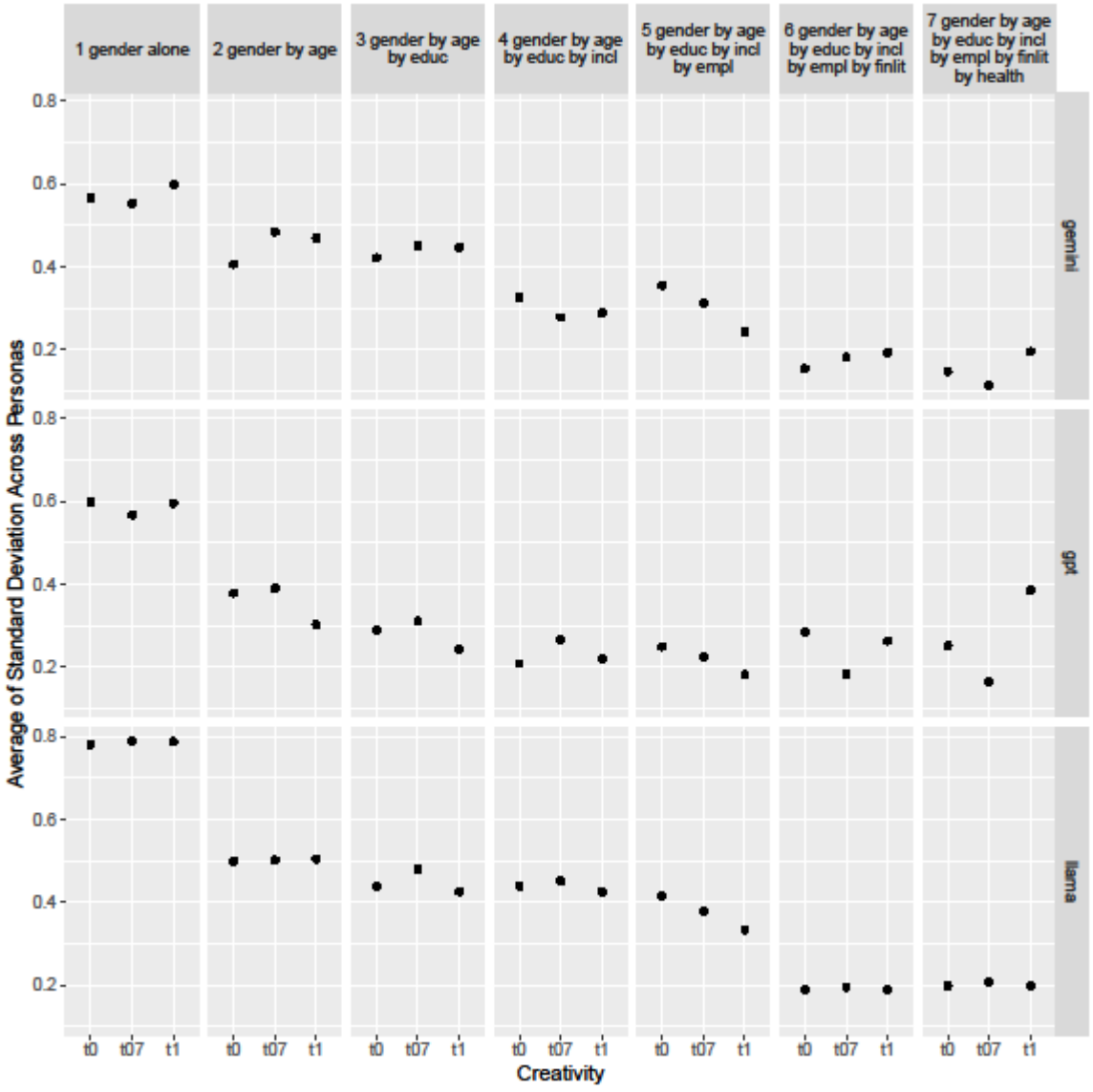


FIGURE 10: New LLM Model

The figure shows the relationship between GPT-4o-mini/GPT-4o and human data.

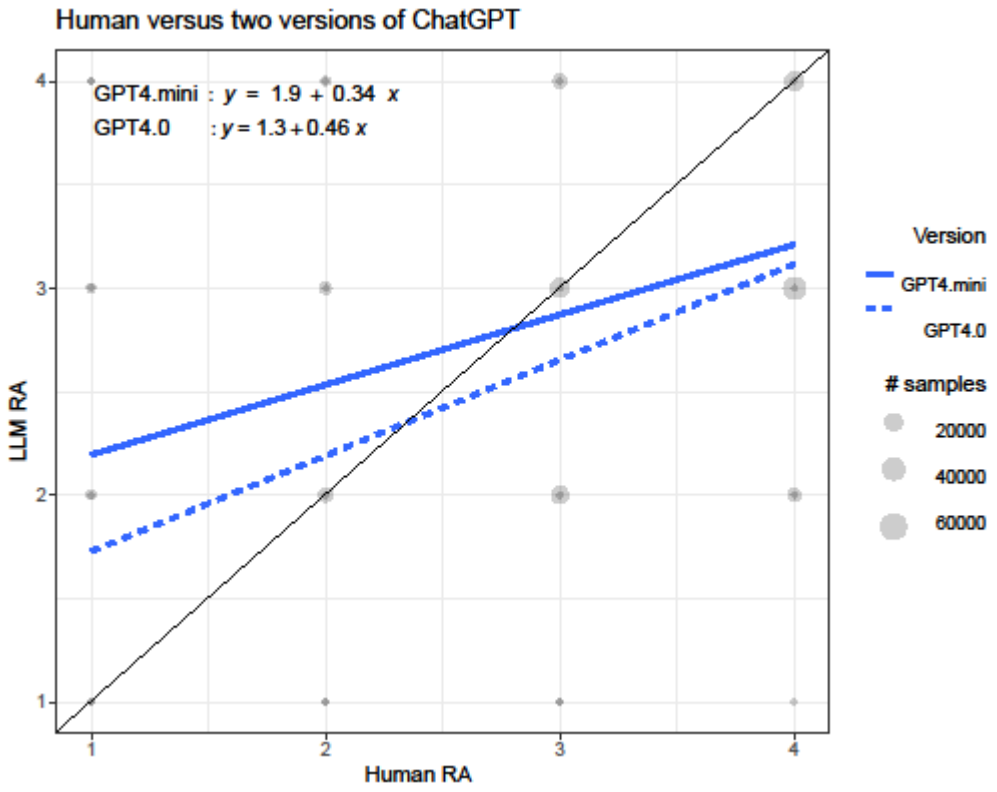


TABLE 1: Summary Statistics

This table presents summary statistics. Table A.1 defines the variables.

				Quantiles		
Variable	N	Mean	S.D.	0.25	Mdn	0.75
GPT RA	149691	2.98	0.72	2	3	3
Human RA	15000	3.33	0.77	3	3	4
GPT RA (ten)	74856	4.84	1.68	4	4	6
Human RA (ten)	7466	3.03	2.67	0	3	5
Female (I)	15000	0.41	0.49	0	0	1
Age (years)	15000	52.2	15.42	40	52	64
Ln(1+Income)	15000	8.05	5.07	0	10.45	11.31
Edu (years)	15000	11.86	2.96	10	12	13
Employed (I)	15000	0.66	0.47	0	1	1
Married (I)	15000	0.88	0.33	1	1	1
FinLit	10000	0.82	0.24	0.67	1	1
Healthy	10000	0.72	0.19	0.6	0.75	0.8

TABLE 2: Correlation Matrix

This table reports pairwise correlation coefficients among the variables. Statistical significance level at 5% is denoted by *. Panel B reports the mean (median) values for the GPT and human risk aversion, and t-statistics (p-value of the Wilcoxon test) of the differences between them.

Panel A: Correlation

	GPT RA	Human RA	Female	Age	Income	Edu	Employed
GPT RA	1						
Human RA	0.3632*	1					
Female	0.2820*	0.2382*	1				
Age	0.3343*	0.0982*	-0.0715*	1			
Income	-0.0777*	-0.0798*	-0.0431*	0.0372*	1		
Edu	-0.2065*	-0.1039*	0.0416*	-0.0840*	0.0105*	1	
Employed	-0.6080*	-0.1844*	-0.1276*	-0.5139*	0.6745*	0.1112*	1
FinLit	0.0329*	-0.0002	0.0816*	-0.0486*	-0.0209*	0.5433*	0.0504*
Married	0.0312*	0.0715*	-0.1559*	-0.0045	0.0073*	0.2860*	0.2447*
Healthy	-0.2719*	-0.1673*	-0.0097*	-0.1492*	0.0296*	0.2527*	-0.003
SCF	-0.3871*	-0.3123*	-0.2517*	0.0401*	0.0751*	-0.4022*	0.1273*
HILDA	0.1372*	-0.0254*	0.1261*	-0.1386*	-0.0360*	0.2819*	-0.0334*
GSOEP	0.2502*	0.3380*	0.1255*	0.0990*	-0.0391*	0.1198*	-0.0943*

	FinLit	Married	Healthy	SCF	HILDA	GSOEP
FinLit	1					
Married	0.4401*	1				
Healthy	0.1479*	0.1240*	1			
SCF	-0.7482*	-0.5314*	-0.0846*	1		
HILDA	0.7467*	0.2652*	0.0843*	-0.4989*	1	
GSOEP	.	0.2660*	.	-0.5006*	-0.4990*	1

Panel B: Mean and Median

Mean GPT RA	Human RA	*t*-stat for (H-G)	Median GPT RA	Human RA	Wilcoxon test *p*-value
2.98	3.33	160.94	3.00	3.00	0.00
149,691	149,691	*N*			
GPT RA (ten)	Human RA (ten)	-210	GPT RA (ten)	Human RA (ten)	
4.84	3.04		4.00	3.00	0.00
74,855	74,855	*N*			

TABLE 3: Baseline Results

The table displays coefficients obtained from OLS regressions. The specification used is as follows:

$$Y_{i(c),t} = \alpha_c + \alpha_t + \beta \times X_{i,t} + \varepsilon_{c,t}$$

where i indexes respondents at a survey year t in a survey c, X_i is a vector of the respondent characteristics used in our persona prompt (age, gender, education, income, employment status, marital status, health status, and financial literacy), and $Y_{i(c),t}$ represents risk aversion for human or GPT. Intercept terms are included but not reported. Standard errors in parentheses are clustered at the respondent level. Table A.1 provides variable definitions. Statistical significance levels are denoted by *, **, and ***, representing significance at the 10%, 5%, and 1% levels, respectively.

	(1) Human RA	(2) GPT RA	Test of the equality	(3) Human RA	(4) GPT RA	Test of the equality
Female	0.215***	0.126***	0.000	0.236***	0.160***	0.000
	(0.01)	(0.01)		(0.02)	(0.01)	
Age	0.002***	0.005***	0.000	0.003***	0.006***	0.000
	(0.00)	(0.00)		(0.00)	(0.00)	
Income	0.001	-0.045***	0.000	0.002	-0.054***	0.000
	(0.00)	(0.00)		(0.00)	(0.00)	
Edu	-0.059***	-0.071***	0.000	-0.062***	-0.049***	0.000
	(0.00)	(0.00)		(0.00)	(0.00)	
Employed	-0.106***	-0.374***	0.000	-0.091***	-0.403***	0.000
	(0.02)	(0.01)		(0.02)	(0.01)	
Married	-0.115***	-0.277***	0.000	-0.032	-0.153***	0.000
	(0.03)	(0.01)		(0.03)	(0.01)	
FinLit				-0.396***	-0.680***	0.000
				(0.04)	(0.02)	
Healthy				-0.387***	-0.403***	0.708
				(0.04)	(0.02)	
Year FE	Yes	Yes		Yes	Yes	
SurveyFE	Yes	Yes		Yes	Yes	
N	15000	149691		10000	99753	
R^2	0.229	0.626		0.185	0.708	

TABLE 4: GPT-Driven Explanation Annotation

The table displays the frequency distribution of Human RA, GPT RA, (used in our baseline results) and GPT RA2 (i.e., GPT-generated RA obtained in the annotation step).

		GPT RA				
Human RA		1	2	3	4	Total
	1	0	65	44	11	120
	2	0	334	109	17	460
	3	1	539	198	40	778
	4	0	132	340	158	630
Total		10	999	720	259	1,988
Human RA = GPT RA						690

		GPT RA2				
Human RA		1	2	3	4	Total
	1	0	60	50	10	120
	2	0	320	110	30	460
	3	10	519	200	49	778
	4	0	100	360	170	630
Total		10	999	720	259	1,988
Human RA = GPT RA						690

		GPT RA				
GPT RA2		1	2	3	4	Total
	1	1	9	0	0	10
	2	0	974	25	0	999
	3	0	86	620	14	720
	4	0	1	46	212	259
Total		10	999	720	259	1,988
Human RA = GPT RA						1807

TABLE 5: Temperature and GPT/Gemini/Llama

This table reports temperature hyperparameter (either 0, 0.7, or 1) and the proportion of LLM-generated risk aversion that are equal to human's by LLM model.

Model	N	% LLM RA = Human RA
Gemini		
t0	98	34.69%
t0.7	98	28.57%
t1	100	31.00%
Total	296	31.42%
GPT		
t0	100	46.00%
t0.7	100	44.00%
t1	100	49.00%
Total	300	46.33%
Llama		
t0	100	42.00%
t0.7	100	42.00%
t1	100	43.00%
Total	300	42.33%
Total		
t0	298	40.94%
t0.7	298	38.26%
t1	300	41.00%
Total	896	40.07%

Appendix

TABLE A.1: Variable definitions

This table defines the variables used in the analysis.

Variable	Definition
GPT RA (ten)	GPT-generated risk aversion level from 1 to 4 (from 0 to 10)
Human RA (ten)	Human risk aversion level from 1 to 4 (from 0 to 10)
Female	=1 if female, zero otherwise
Age (years)	=age of respondent
Ln(1+Income)	=Ln(1+income (in USD, AUD, or Euro)) in a year
Edu	= education attainment (in years)
Employed	=1 if employed, zero otherwise
Married	=1 if married, zero otherwise
FinLit	=normalized financial literacy by dividing it by 5 or 3: Financial literacy is measured using the number of correct answers on the five [HILDA] (or three [SCF]) financial literacy questions.
Healthy	=normalized health status by dividing it by 4 or 5: Health status is measured using excellent, fair, good, or poor in SCF [or very good in HILDA].
Year	Survey year
Keywords in Explanation	'gender': ['male', 'female', 'gender', 'sex', 'woman', 'man'], 'age': ['age', 'old', 'young', 'senior', 'child', 'teenager', 'adult'], 'income': ['income', 'salary', 'wealth', 'poverty', 'rich', 'poor', 'financial'], 'education': ['education', 'school', 'college', 'university', 'degree', 'diploma', 'literate'], 'employment_status': ['employed', 'unemployed', 'job', 'work', 'occupation', 'career'],'marital_status': ['married', 'single', 'divorced', 'widowed', 'relationship', 'spouse'], 'health_status': ['health', 'sick', 'ill', 'disease', 'condition', 'healthy', 'wellbeing'], 'country_of_living': ['country', 'nation', 'residence', 'citizen', 'location', 'living'], 'financial_literacy': ['financial literacy', 'financial knowledge', 'money management','budgeting', 'investing'], 'cognitive_ability': ['cognitive', 'intelligence', 'thinking', 'reasoning', 'memory', 'learning', 'ability']