Kyoto University,
Graduate School of Economics
Discussion Paper Series

Comparing Risk Preferences and Reference Dependence in Humans and AI: A Persona-Based Approach with Fine-Tuning

Ryota Iwamoto, Takunori Ishihara, and Takanori Ida

Discussion Paper No. E-25-006-V2

Graduate School of Economics
Kyoto University
Yoshida-Hommachi, Sakyo-ku
Kyoto City, 606-8501, Japan

First issued: July 2025
Revised: July 2026

Comparing Risk Preferences and Reference Dependence in Humans and AI:
A Persona-Based Approach with Fine-Tuning

Ryota Iwamoto[1], Takunori Ishihara[2], and Takanori Ida[3]

Abstract:

This study empirically investigates the differences in risk preferences and reference dependence between humans and generative AI. We conduct a nationwide online survey of 4,838 individuals and generate AI responses under identical conditions by using personas constructed from demographic attributes. The results show that in gain domains, both humans and the AI select risk-averse options and exhibit similar preference patterns. However, in loss domains, AI shows a stronger risk-loving tendency and responds more sharply to individual attributes such as gender, age, and income. We retrain the AI by fine-tuning it based on human choice data. After fine-tuning, the AI's risk preference tendency moves closer to that of humans, with loss-related decisions showing the greatest improvement. Using Brier score and log loss, we suggest that fine-tuning partly reduces the gap in risk preference and reference dependence between AI and humans.

JEL Classification: D91, C91

Keywords: bias, risk preference, reference dependence, generative AI, persona, fine-tuning

[1] Graduate School of Economics, Kyoto University, iwamoto.ryota.82a@st.kyoto-u.ac.jp

[2] Institute for International Academic Research, Kyoto University of Advanced Science, ishihara.takunori@kuas.ac.jp

[3] Graduate School of Economics, Kyoto University, ida@econ.kyoto-u.ac.jp

1. Introduction

As artificial intelligence (AI) has increasingly automated judgment processes, concerns have arisen regarding its legitimacy and explainability. Human-in-the-loop (HITL), a framework that integrates human and AI in economic decision making (Rahwan, 2018; Rahwan et al., 2019), offers both theoretical and practical responses to these concerns. The comparison of bias structures between humans and AI plays a central role in understanding the value of HITL. AI often inherits biases from its training data. It tends to reproduce the structural biases embedded in past economic behavior and institutions. Meanwhile, as behavioral economics reveals, human decision making also reflects systematic deviations due to emotion-driven cognitive biases. Understanding these distinct bias structures and creating designs that allow mutual correction between humans and AI are essential for achieving rationality. Therefore, when designing HITL-based decision systems, the commonalities and differences between algorithmic and cognitive biases, from both theoretical and empirical perspectives, must be examined to provide a theoretical foundation for developing reliable decision-support systems that leverage the complementary strengths of human judgment and AI reasoning.

The role of generative AI as a decision-support tool has also attracted growing attention in the context of policymaking. Generative AI may contribute to the policy formation process through functions such as summarizing and generating policy-relevant information, as well as supporting decision-making. At the same time, concerns have been raised regarding the potential generation and diffusion of misinformation, the reproduction of biases against particular social groups, and even the amplification of existing socioeconomic inequalities. These considerations suggest that both the benefits and risks of policy applications must be carefully evaluated (Capraro et al., 2024). In addition, the effective use of generative AI as a decision-support tool requires an understanding of the ways in which biases may arise systematically in human decision-making.

To advance this line of analysis, systematic biases that arise in human decision-making must be first clarified. Humans often deviate from the predictions of standard utility-maximization models. In particular, under conditions of risk or uncertainty, existing studies observe behavioral patterns such as loss aversion. Kahneman and Tversky (1979) organize these tendencies into a theoretical framework known as prospect theory, which is the basis for studies that document various forms of irrational judgments. Tversky and Kahneman (1992) estimate the degree of loss aversion at approximately $\lambda = 2.25$ and suggest that people experience losses about twice as intensely as equivalent gains and therefore tend to avoid them. Falk et al. (2018) conducts a cross-national survey of approximately 80,000 individuals in 76 countries. They find that loss aversion varies not only across cultures but also according to individual attributes, such as age, gender, and education. Brown et al. (2024) conduct a meta-analysis of 150 studies in economics and psychology, reporting a median loss aversion coefficient of approximately $\lambda = 1.69$. These findings offer a foundational understanding of human behavior that

informs the design of HITL systems, particularly regarding the role of human agents within such frameworks.

In recent years, the rise of generative AI, exemplified by models such as ChatGPT, has sparked growing interest in whether the cognitive biases commonly observed in humans also appear in AI. As generative AI models are trained on large-scale human-generated language data, these models may inherently acquire human-like tendencies and biases. If generative AI can replicate such biases, it can serve as a practical alternative in fields such as social surveys, economic experiments, and policy simulations. However, concerns remain regarding the amplification of biases and the unpredictability of AI-generated outputs. Thus, understanding the behavioral patterns and limitations of AI responses has become a central research challenge. Against this backdrop, more studies have compared the judgment patterns of humans and AI to identify similarities and differences.

A key question is whether assigning specific personas to generative AI results in more human-like outputs. A persona refers to a set of hypothetical individual attributes, such as age, gender, education, income, and cultural background, assigned to an AI agent to generate responses with greater consistency and personality. Many previous studies embed such attribute information directly into the prompt, reporting improved output coherence. Park et al. (2024) use detailed interview data from 1,052 American respondents to test how well AI agents replicate actual human behavior when assigned attributes such as age and political orientation. Their results indicated a matching rate of up to 85%. Jia et al. (2024) assign demographic profiles to GPT models and measure their risk preferences and loss aversion. They find that differences in gender and educational attainment significantly affect the estimated loss aversion coefficients and probability weighting. In summary, the findings suggest that generative AI can replicate human-like behavioral patterns and provide valuable insights for future empirical and applied research.

Several studies reveal that generative AI shows consistent patterns in gain-and-loss scenarios. A common observation is that generative AI exhibits risk preferences that partially differ from those of humans. Qiu et al. (2024) report that GPT-4 has a loss aversion coefficient of 1.09, compared to 2.56 for actual consumers, suggesting that the model underestimates losses relative to gains. Jia et al. (2024) find that although generative AI generally behaves in a risk-averse manner, it tends to outweigh the probabilities of rare events. Meanwhile, model-level comparisons reveal several important differences. Ross et al. (2024) demonstrate that GPT-4 generates the most stable and human-like preference structures compared to other models, such as Claude and Gemini, which exhibit greater variability in their outputs. Chen et al. (2025) highlight the sensitivity of AI preferences to phrasing prompts. Even when presented with identical choice tasks, generative AI produces different responses depending on contextual cues, suggesting a lack of internal consistency and a high degree of dependence on prompt design.

Fine-tuning is the process of retraining a pre-trained foundation model on additional data related to

specific tasks or domains, thereby adjusting its output tendencies and performance to meet objectives. Although many existing studies do not explicitly fine-tune models using human data, they are concerned that reinforcement learning from human feedback (RLHF), often implemented as part of HITL, may significantly affect the behavior of the model. Santurkar et al. (2023) show that large language models (LLMs) tend to produce outputs biased toward liberal, highly educated, and high-income groups, and underrepresent the views of older adults, low-income individuals, and religious communities. The RLHF process may reinforce specific value systems and amplify the underlying biases. Hagendorff et al. (2023) note that although RLHF improves the models' ability to avoid intuitive errors and generate accurate responses, it risks entrenching normative assumptions within the model. In contrast, Park et al. (2024) adopt a different approach by embedding actual human response data into a prompt to construct individual AI agents. Although this does not constitute formal fine-tuning, it effectively injects human knowledge into the model and functions as an implicit calibration method. These findings suggest that generative AI exhibits human-like behavioral tendencies when conditioned on social attributes and contextual framing but still lacks consistent judgment across various settings. Although RLHF may enhance coherence and response quality, empirical evaluations of its effects remain limited. Therefore, assessing how well generative AI replicates human decision-making biases before and after fine-tuning represents a key research agenda in both behavioral economics and applied AI.

This study focuses on reference dependence, which is closely related to diminishing sensitivity in risk preferences and loss aversion. It aims to empirically examine the similarities and differences in decision-making biases between humans and generative AI (GPT-4). Specifically, it analyzes how individual attributes influence these biases across the two entities. Our study designs a unified experimental framework to analyze the presence and direction of biases using equivalent tasks for both humans and AI. We conduct a nationwide online survey of 4,838 Japanese adults and explore the relationship between individual attributes such as age, gender, household income, and their response patterns. Based on these demographic profiles, we construct virtual personas and present the same set of questions to the GPT-4 under multiple temperature settings. We then quantitatively compare the AI-generated responses with human preferences to evaluate both the divergence and degree of alignment in their underlying decision structures.

At this point, *reference dependence* is observed as risk-averse behavior in the gain domain and risk-seeking behavior in the loss domain, using zero yen as the reference point[4]. Our empirical analysis reveals that, in the gain domain, humans and GPT-4o exhibit broadly similar patterns of risk aversion. In contrast, in the loss domain, GPT-4o demonstrates significantly stronger risk-seeking tendencies than

[4] This study does not directly estimate the differences in the value function between the gain and loss domains. Rather, it focuses on the key behavioral implication of loss aversion and diminishing sensitivity—namely, the reference dependence of risk preferences across the two domains—which is a phenomenon observed in both humans and AI agents.

humans, indicating a divergence in risk attitudes that may stem from differences in the degree or functional form of reference dependence. Moreover, the results suggest that GPT-4o is more sensitive to demographic cues in its decision-making, with individual attributes such as age and income exerting a stronger influence on AI-generated responses than on human choices.

We also fine-tune the AI model using web-based survey data on human reference dependence. Our results show that the AI's tendencies in terms of patterns of risk preference and the marginal effects for demographic attributes become closer to human tendencies. However, in both the gain and loss domains, these tendencies appear to exhibit a slight overshoot. We also use the Brier score and log loss as a supplementary assessment of similarity. Fine-tuned AI exhibits a shift in a direction closer to the human choice distribution with respect to risk preference and reference dependence.

Moreover, this study can be positioned as an example of research that leverages human knowledge for the evaluation and additional training of machine learning models, insofar as it uses human response data to assess and adjust the response tendencies of generative AI. However, this study does not take the design of human-in-the-loop systems, in which humans and machines interact jointly in decision-making, as its object of analysis. Rather, its primary focus is to examine the validity and limitations of generative AI as synthetic subjects by presenting identical tasks to both humans and generative AI and comparing the similarities and differences in their responses. In this sense, the study is positioned as an attempt to examine the extent to which established behavioral economic patterns of risk preference and reference dependence are reproduced in generative AI and how these patterns may change through additional training based on observed human choice data.

The remainder of this paper is organized as follows. Section 2 reviews previous studies that compare generative AI with human behavior. Section 3 describes the survey design and data collection procedures. Section 4 introduces the fine-tuning process based on human choice patterns. Section 5 describes the estimation methods for behavioral biases and compares the risk preference tendencies of humans, pre-trained AI, and fine-tuned AI and summarizes their interrelationships. Section 6 discusses the main findings. Lastly, Section 7 concludes the study and outlines future research directions.

2. Related Literature

We review previous studies that compare decision-making patterns and cognitive biases between generative AI and humans[5]. Table 1 lists the studies based on the following seven dimensions: 1) type

[5] Several studies examine the research applications of LLMs, outlining both their potential and limitations. Demszky et al. (2023) provide a systematic overview of how LLMs can be applied in psychology and highlight the risks of biased outputs when cultural biases or underrepresented groups are not adequately captured in the training data. Hagendorff (2023) introduces the concept of "machine

of bias or experimental task examined; 2) version and type of generative AI used; 3) target population and sample size; 4) extent to which LLMs replicate average human behavior; 5) replicability of human responses when conditioned on socioeconomic attributes; 6) whether the model incorporates human feedback, such as fine-tuning; and 7) metrics used to evaluate differences in response distributions between humans and LLMs.

＜Table 1: Prior Studies on Decision-Making and Cognitive Biases: Generative AI vs. Humans＞

We introduce studies that compare human decision-making with that of large LLMs through economic experiments. We first focus on game-theoretic experiments that involve social preferences such as altruism and strategic interactions. Although these studies do not directly compare AI behavior with human data, they evaluate LLM behavior in economic games by allowing multiple models to interact with one another (e.g., Akata et al., 2025; Guo, 2023; Kitadai et al., 2024; Tsuchihashi, 2023; Lorè and Heydari, 2024). Akata et al. (2025) conduct repeated games using GPT-3, GPT-3.5, and GPT-4 in scenarios such as the Prisoner's Dilemma and the Battle of the Sexes, each with different payoff matrices. They show that GPT-4 exhibits human-like cooperative behavior but frequently struggles to maintain implicit coordination. However, coordination success rates significantly improve when a "cooperative persona" is introduced in the prompt. Guo (2023) assigns GPT-4 personas such as "fair" and "selfish" and repeatedly runs ultimatum games and the Prisoner's Dilemma. They show that the "fair" persona generates offer and rejection thresholds close to human averages, while the "selfish" persona leads to low offers and high acceptance rates. Kitadai et al. (2024) conduct simulations of the ultimatum game using GPT-3.5 and GPT-4. They demonstrate that enhancing the reasoning abilities of GPT-based generative agents yields outcomes closer to theoretical predictions than to actual human experimental data. Lorè and Heydari (2024) present four two-player games under five contextual framing conditions and compare the behavior of GPT-3.5, GPT-4, and LLaMa-2. They reveal that GPT-4 tends to select strategies consistent with game-theoretic rationality depending on the game structure.

---

psychology," proposing a framework that treats LLMs as virtual participants in psychological experiments and draws attention to the ethical concerns associated with this approach. Sarstedt et al. (2024) review studies comparing so-called "silicon samples" with human respondents to assess the applicability of LLMs in consumer behavior and marketing research. They find that LLMs can achieve high replicability in tasks such as brand evaluation and framing effects but also note their limitations in replicating preference formation and behavioral decision-making tasks. These reviews offer valuable guidance for incorporating LLMs into social science research and provide critical perspectives on ethical governance and future application areas.

In contrast, GPT-3.5 and LLaMa-2 exhibit greater sensitivity to contextual framing. Tsuchihashi (2023) examines GPT-3.5's bidding behavior in sealed-bid auctions. They find that in first-price auctions (FPA), the model overbids similarly to humans, whereas in second-price auctions (SPA), it submits truthful bids in line with theory. Furthermore, when given a "student persona," GPT-3.5 places more theoretically consistent bids in FPA and underbids in SPA, suggesting that persona conditioning affects strategic behavior.

Several studies compare the behavior of LLMs with human decisions through experiments on social preferences (e.g., Horton, 2023; Brookins and DeBacker, 2024; Mei et al., 2024; Xie et al., 2024; Capraro et al., 2025). Horton (2023) conceptualizes LLMs as Homo Silicus, a virtual economic agent, and presents GPT-3 with standard behavioral economics tasks, such as fairness evaluations, status quo bias, and responses to minimum wage policies. They show that the model exhibits biases consistent with those observed in previous human experiments. Moreover, its decision patterns vary depending on the assigned persona, such as selfish, fair-minded, or efficiency-oriented. Brookins and DeBacker (2024) repeatedly run dictator and prisoner's dilemma games using GPT-3.5 to assess its tendency toward fairness and cooperation. They find that the model tends to behave more altruistically than humans, particularly in contexts that emphasize efficiency, in which its behavior aligned more closely with human patterns. Mei et al. (2024) analyze how GPT-3.5 and GPT-4 replicate human behavior by comparing model outputs with a dataset of approximately 90,000 human decisions across six game types. Using a Turing test framework, they find that GPT-4's responses closely resemble the human average. However, the model also displays a centralizing tendency, avoiding extreme responses. Xie et al. (2024) evaluate five models, including GPT-4o, using games such as the dictator and ultimatum game. They assess the differences in response distributions between LLMs and humans using the Wasserstein distance. Overall, LLMs demonstrate fairer and more cooperative behavior than humans, with GPT-4o producing the distribution most similar to human responses. Capraro et al. (2025) used models such as GPT-4 and GPT-4o to predict human donation behavior distributions based on 106 dictator game experiments previously conducted across 12 countries and evaluated performance using mean and distributional errors. While generative AI tended to overpredict average donation amounts, it nonetheless exhibited some ability to reproduce the shape of the donation distribution.

Meanwhile, several experimental studies examine loss aversion and risk preferences in LLMs (Chen et al., 2023; Jia et al., 2024; Qiu et al., 2024; Ross et al., 2024; Chen et al., 2025; Macmillan-Scott and Musolesi, 2024). Chen et al. (2023) conducted decision-making experiments using GPT-3.5 in the domains of risk preference, time preference, social preference, and food choice and evaluated the rationality of human and generative AI decision-making using rationality measures based on the generalized axiom of revealed preference (GARP). Their results suggest that GPT tends to exhibit higher rationality than humans while being sensitive to problem framing. Jia et al. (2024) present choice tasks related to loss aversion and probability weighting to GPT-4, Claude, and Gemini. They show that

all models exhibit risk-averse behavior, with GPT-4 demonstrating loss-aversion patterns similar to those of humans. They also report substantial variations in output depending on the prompt conditions based on demographic attributes. Ross et al. (2024) comprehensively analyze 12 LLMs, including GPT-4, focusing on biases such as inequality aversion, loss aversion, and time discounting. The models reveal intense guilt and weak envy in terms of social preferences. While they respond rationally to gains, their behavior under losses deviates from rational expectations. Additionally, the LLMs display stronger time discounting than human subjects. Chen et al. (2025) examine 18 cognitive biases, including risk attitudes. They show that GPT-3.5 tends to be risk-seeking for gains and risk-averse for losses, while GPT-4 exhibits more consistent risk aversion and lower susceptibility to framing effects. Macmillan-Scott and Musolesi (2024) evaluate 12 cognitive bias tasks, including the conjunction and gambler's fallacies. They show that GPT-3.5 reproduces many human-like biases, whereas GPT-4 achieves a higher overall accuracy. Finally, although not based on an economic experiment, Qiu et al. (2024) simulate health insurance plan choices (n = 5,998) using GPT-4 to assess whether the model could replicate human decisions under risk. They show that while the aggregate choice distributions resemble those of humans, individual-level agreement is low. The estimated loss aversion coefficient for GPT-4 (1.09) is substantially lower than the human average (2.56), suggesting that the model systematically underweights losses relative to human decision-makers.

Other studies examine the reproducibility of survey responses using generative AI. Santurkar et al. (2023), Park et al. (2024), Dominguez-Olmedo et al. (2025), and Toubia et al. (2025) compare AI-generated outputs with actual public opinion and survey data to evaluate the degree of alignment and divergence. In the domain of cognitive biases, some studies investigate the extent to which LLMs replicate cognitive errors and reflective thinking patterns (Aher et al., 2023; Binz and Schulz, 2023; Hagendorff et al., 2023; Wang et al., 2025). Additionally, research on political attitudes and voting behavior analyze the reproducibility of AI outputs under specific political ideologies and demographic conditions (Argyle et al., 2023; Motoki et al., 2024; Bisbee et al., 2024). Lastly, additional studies explore the behavioral alignment between AI and humans in various domains, covering topics such as market research, economic forecasting, environmental awareness, and tourism behavior (Brand et al., 2023; Bybee, 2023; Lee et al., 2024; Li et al., 2024; Xiong et al., 2024).

These prior studies suggest that generative AI may reproduce human decision-making to some extent under particular conditions; however, several limitations remain in the methods of comparison and evaluation frameworks employed in prior studies. First, many studies do not directly compare human and generative AI behavior under identical experimental tasks, raising the possibility that differences in question design or sample composition may affect interpretation of the results. Second, relatively few studies systematically incorporate demographic attributes and quantitatively evaluate differences in response tendencies and attribute-specific effects under such conditions. Third, studies that explicitly implement fine-tuning based on observed human data and assess its effects from the perspective of out-

of-sample performance remain extremely limited.

Accordingly, this study addresses these gaps in three ways. First, using identical choice tasks, we directly compare responses from a large-scale web survey of human participants with responses generated by generative AI to examine their differences in risk preference and reference dependence. Second, we assign the generative AI with personas reflecting demographic attributes based on observed survey data, thereby enabling the evaluation of not only average tendencies but also the differences in response tendencies under matched attribute conditions. Third, we fine-tune generative AI using observed human choice data and evaluate the effects of this procedure from the perspective of out-of-sample performance using metrics such as the Brier score and log loss.

3. Survey Design

3.1. Design of the Web-Based Survey

We conduct a web-based survey (December 2024) through an online survey company MyVoice (https://www.myvoice.co.jp/), targeting domestic residents aged 20–65. We divide Japan into nine regions and allocate respondents to balance gender and age groups within each region. Participants are informed in advance that the survey is an academic study aimed at examining the relationship between consumer attributes and preferences and are notified in advance that compensation would be provided for participation. Specifically, respondents who complete all survey items receive a uniform payment of JPY 75. We obtain responses from 5,040 participants. This survey does not include explicit attention-check items. Instead, as an ex-post exclusion criterion, respondents who do not complete the survey are excluded from the analysis (n = 202). The final sample includes 4,838 participants. Although regional quota sampling is employed in this study, no ex post regional weighting is applied in the estimation. The survey includes individual attributes, such as gender, age, and household income. It also contains multiple questions on cognitive biases and psychological characteristics that cover Big Five personality traits, time preferences, ultimatum games, and the trolley problem. In this study, following careful preliminary analysis, we use gender, age, and household income as individual attributes, as these variables have been widely employed in prior research and exhibit statistically significant associations in the data used in this study[6]. To maintain a clear analytical focus, we concentrate on risk preference and reference dependence grounded in prospect theory when examining psychological characteristics. Table 2 presents the descriptive statistics for the respondents' gender, age, and household income[7]. With

[6] For the analysis results using educational attainment, occupation, and Big Five personality traits, see Appendix A.1.

[7] Income is categorized into six groups: "Less than JPY 3 million," "JPY 3 million to less than JPY 5 million," "JPY 5 million to less than JPY 7 million," "JPY 7 million to less than JPY 10 million," "JPY 10 million to less than JPY 15 million," and "JPY 15 million or more." Class values are set as follows: JPY 3 million, JPY 4 million, JPY 6 million, JPY 8.5 million, JPY 12.5 million, and JPY 15 million.

respect to nationality, it is not included in the analysis because the survey is conducted among residents of Japan and the available information for identifying sufficient heterogeneity within the sample is too limited to meaningfully assess nationality-based differences.

＜Table 2: Descriptive Statistics＞

3.2. Questions on Prospect Theory

In this section, we discuss the questionnaire used to measure risk preferences and reference dependence. As a reference, the following choice tasks concerning gains and losses are presented to students at Stanford University and the University of British Columbia by Tversky and Kahneman (1988):

Problem 1. (n=126)

Assume yourself richer by $300 than you are today. You are offered a choice between

A. A sure gain of $100, or

B. A 50% chance to gain $200 and A 50% chance to lose $0.

Problem 2. (n=128)

Assume yourself richer by $500 than you are today. You are offered a choice between

A. A sure loss of $100, or

B. A 50% chance to lose $200, and A 50% chance to lose $0.

Participants are presented with two choices (A and B) with equivalent expected values in both the gain and loss domains. In Problem 1, which corresponds to the gain domain, 72% of participants select the sure gain option (A), whereas 28% choose the risky option (B). This indicates a risk-averse preference in the gain domain, in which most participants prefer certainty over risk. However, in Problem 2, which corresponds to the loss domain, 36% of the participants select the sure option (A), while 64% choose the risky option (B). This suggests a risk-seeking tendency in the loss domain, in which participants gamble on the possibility of avoiding loss by accepting a higher potential loss.

Another important point is the difference in the choice tendencies between Problems 1 and 2. Although both problems present identical monetary amounts and probability structures from the same wealth reference point, the observed choices differ. This preference asymmetry reflects the existence of reference dependence—a psychological tendency in which individuals are risk-averse in the gain domain but risk-seeking in the loss domain[8]. This tendency highlights a behavioral pattern that deviates

[8] Kahneman and Tversky (1979) reports a similar tendency. The structure of the questions and the results in their study are as follows. In PROBLEM 1, the proportions of participants choosing options

from the rational decision-making predicted by the standard expected utility theory, showing that human decisions are heavily influenced by frameworks such as reference points. Based on Tversky and Kahneman (1988), we present the following questions:

Question 1. Imagine you receive an additional JPY 30,000 on top of your current wealth and are asked to choose between the options below. Which option would you choose?

Option 1: Receive a guaranteed JPY 10,000

Option 2: A 50% chance of receiving 20,000 yen and a 50% chance of receiving nothing

Question 2. Imagine you receive an additional JPY 50,000 on top of your current wealth and are asked to choose between the options below. Which option would you choose?

Option 1: Lose 10,000 yen with certainty

Option 2: A 50% chance of losing JPY 20,000 and a 50% chance of losing nothing

Since our survey targets domestic residents in Japan, questions are presented in the Japanese language, and monetary amounts are displayed in JPY (with an exchange rate of USD 1 = JPY 100).

3.3. Survey Design for Generative AI

We implement "personas" constructed from actual web-based survey attribute information into generative AI and compare the AI's decision tendencies with those of humans under identical conditions. Specifically, we extract representative patterns of attributes from human data and input them into prompts to assign virtual personas to the AI. Then, we present the same questions as in the web-based survey to these personas and collect their responses, enabling a condition-controlled comparative analysis. We use OpenAI's GPT-4o as the generative AI model. Unlike the general ChatGPT interface, GPT-4o is accessible via the OpenAI API, allowing program-based operation. Using the API, we can automate large-scale response generation with Python code, enabling systematic analysis under numerous persona conditions[9]. For details regarding the dataset, prompts, and Python scripts used in

---

A and B are 16% and 84%. In contrast, in PROBLEM 2, the proportions of participants choosing options C and D are 69% and 31%, respectively.

Problem 1.

In addition to whatever you own, you have been given 1,000. You are now asked to choose between: A: ($1,000, 0.50), and B: ($500).

Problem 2.

In addition to whatever you own, you have been given 2,000. You are now asked to choose between: C: ($-1,000, 0.50), and D: ($-500).

[9] In addition to GPT-4o, we also collect responses to the same questions from other chat-based language models, including OpenAI's GPT-3.5, Google's Gemini 2.5 Flash, and DeepSeek's DeepSeek-V3. The results of these comparisons are discussed in Section 6.

this study, see Appendix A.2.

To control variability in the generated responses, we also adjust the "temperature" parameter, which controls the randomness, and collect responses under three different temperature settings[10]. By adjusting the temperature, we could evaluate the impact of probabilistic fluctuations on AI decision tendencies[11].

Through these procedures, we reproduce 4,838 personas with attributes identical to those of the web-based survey participants on GPT-4o and obtain responses to the two-choice tasks (see Table A.2.2)[12]. Then, we apply a logit model to the response data and quantitatively analyze the generative AI's tendencies regarding risk preferences and reference dependence.

3.4. Estimation Methods for Behavioral Biases

We present the estimation method used to clarify the impact of respondent attributes on risk preferences and reference dependence[13]. In the questionnaire, the respondents are asked to choose between two options: "Option 1: JPY 10,000 with 100%" or "Option 2: JPY 20,000 with 50%." Therefore, the responses can be treated as binary variables, and using a logit model is appropriate. The choice probability in the logit model is as follows:

$$P(Y_i = 1 | X_{1i}, X_{2i}, X_{3i}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i})}}$$

where i denotes each respondent. The dependent variable Y_i equals 1 if the respondent selects "Option 1: JPY 10,000 with 100%" and 0 if they select "Option 2: JPY 20,000 with 50%." X_{1i} is a dummy variable for the gender of respondent i, taking the value of 1 if the respondent is female and 0 if the respondent is male. X_{2i} represents the respondent i's age, scaled in 10-year units (per decade). X_{3i} indicates the respondent i's household income, scaled in units of JPY 1 million.

4. Fine-Tuning

This study also aims to replicate human behavioral biases within generative AI by training it on real

---

[10] The temperature parameter ranges from 0 (minimum) to 2 (maximum), with a default value of 1. Higher temperatures lead to increased randomness in the generated text. However, when the temperature exceeds 1, the analysis results become unstable in our setting.

[11] In the main analysis of this study, the temperature parameter is fixed at its default value of 1. For analyses using alternative temperature settings (0 and 0.5), see Appendix A.3.

[12] Unlike the general ChatGPT interface, responses generated via the OpenAI API are stateless and independent, ensuring response independence comparable to that in web-based surveys.

[13] From this point forward, we refer to both the web-based survey participants and generative AIs with implemented personas as "respondents." The terms "individual" and "answer" are used in the same unified manner.

human response data and to evaluate the proximity between humans and AI before and after training. Thus, we employ fine-tuning as the training method for the generative AI. Fine-tuning refers to the retraining of a pre-trained model, originally trained on large-scale data, on a new dataset to specialize in a specific task or domain (Devlin et al., 2019). In the case of OpenAI's GPT model, fine-tuning is performed by preparing a dataset in the JSONL format, uploading it to the OpenAI API system, and initiating the training process.

In this study, we use the data constructed from the actual response values provided by individual respondents in the web-based survey as the training data. After completing the fine-tuning process, we submit to the fine-tuned model the same questions as those for the human participants and pre-trained AI. We estimate the predicted probability using a logit model to clarify the differences in behavioral tendencies among the three groups. The details of the dataset and Python scripts used for fine-tuning are provided in Appendix A.4.

In the fine-tuning process, determining the training methods and parameter settings in advance is necessary. For the training methods, we adopt an attribute-group-level hold-out approach informed by the logic of group K-fold cross-validation. Specifically, the 4,838 samples are partitioned into eight groups based on three attributes: gender, age, and income (Table A.4). For each iteration, data from seven groups excluding a given $group_i$ ($1 \leq i \leq 8$), denoted as $group_{-i}$, are used as training data, consisting of respondents' gender, age, household income, and response values, to fine-tune the AI and construct a fine-tuned model, AI_{-i}. The fine-tuned AI_{-i} is then prompted using personas replicating the attributes (gender, age, and household income) of individuals in the *held-out* $group_i$, and responses are obtained accordingly. This process is repeated for i = 1, …, 8, yielding fine-tuned AI responses corresponding to all 4,838 personas. This design ensures that human responses and corresponding persona-based AI responses within the same attribute group are never included in the training and test data simultaneously. Consequently, changes in proximity observed after fine-tuning can be evaluated as out-of-sample performance with respect to the attribute groups excluded from training.

For the parameters, we consider epoch and learning rate multiplier (LR). Epoch is the number of training iterations. The higher the value, the more the model learns; in this case, the risk of overfitting increases. LR is used to adjust the learning speed. A higher value leads to faster and more efficient learning and easier convergence; however, it also increases the risk of overfitting.

We conduct preliminary fine-tuning using a randomly selected subsample of 500 respondents to select the appropriate learning method and parameter settings. Then, we instruct the fine-tuned model to generate responses to the two questions and evaluate them using logit analysis. Based on these preliminary results, we adopt epoch = 3 and LR = 0.02 for Question 1 and epoch = 1 and LR = 0.2 for Question 2.

We apply the same procedure as the one for the pre-trained AI (Section 3.3) to the fine-tuned AI. Specifically, we reproduce 4,838 personas on the generative AI, each with the same attributes as those

in the web-based survey results, and obtained responses to two-choice questions. The resulting responses are analyzed using the logit model to examine the tendencies in risk preferences and reference dependence.

5. Comparison of Biases between Humans and AI

5.1. Overview of Choice Patterns

Table 3 presents the choice results for each question obtained from humans, pre-trained AI and fine-tuned AI. For each question, a binary variable is defined: Option 1 (a certain gain/loss of 10,000 yen) is coded as 1 and Option 2 as 0. The average of this variable represents the average selection rate of Option 1.

<Table 3: Choice Results: Human vs. Pre-trained AI>

We first review the results of the web-based survey. For Question 1, 88.4% of the human respondents select Option 1 (a sure gain of JPY 10,000), whereas only 11.6% choose Option 2 (a 50% chance of gaining JPY 20,000 or nothing). This result suggests a strong risk-averse tendency in the gain domain. For Question 2, 57.7% choose Option 1 (a sure loss of JPY 10,000), while 42.3% choose Option 2 (a 50% chance of losing JPY 20,000 or nothing). Although the result indicates some degree of risk-seeking behavior, no clear tendency is found in the loss domain.

Then, we examine the response results from the pre-trained AI. For Question 1, 91.1% of the respondents select the sure gain option, indicating a clear tendency toward risk aversion in the gain domain. However, for Question 2, only 11.0% choose the sure loss option, whereas the majority opt for the probabilistic loss, indicating a distinct risk-seeking tendency in the loss domain. In summary, the results suggest that humans and pre-trained AI display a high level of risk aversion in the gain domain. In contrast, in the loss domain, while the pre-trained AI exhibits a clear risk-seeking tendency, the human responses do not display a similarly strong pattern. This discrepancy suggests that preferences in the loss domain may differ between humans and pre-trained AI. When we compare the selection rates for Option 1 in Questions 1 and 2, pre-trained AI results are 91.1% and 11.0%, indicating stronger reference dependence. For humans, the rates are 88.4% and 57.7%, reflecting relatively weaker reference dependence compared to pre-trained AI.

Next, we examine the response results from fine-tuned AI. For Question 1, respondents choosing Option 1 (94.2%) significantly exceed those who choose Option 2 (5.8%). This suggests a strong tendency toward risk aversion in the gain domain. For Question 2, more respondents select Option 1 (62.1%) over Option 2 (37.9%). This indicates a weak tendency toward risk-seeking in the loss domain.

Moreover, we compare the results of humans, pre-trained AI, and fine-tuned AI. For Question 1, the

fine-tuned AI's selection rate (94.2%) is higher than that of the pre-trained AI (91.1%), which is higher than the human average (88.4%). This suggests overfitting in the form of excessive risk aversion in the gain domain. Similarly, for Question 2, the fine-tuned AI's average selection rate (62.1%) is closer to that of humans (57.7%), indicating that it appropriately learns human tendencies toward risk-seeking in the loss domain.

Comparing the selection rates for Option 1 between Questions 1 and 2, the fine-tuned AI rates for Questions 1 and 2 (94.2% and 62.1%, respectively) are higher than those of the pre-trained AI (91.1% and 11.0%, respectively), indicating that the fine-tuned AI has a weaker reference dependence and a pattern similar to that of humans (88.4% and 57.7%, respectively).

5.2. Analysis Based on Marginal Effects

Tables 4 and 5 present the estimated marginal effects and average predicted probabilities derived from the logit model using datasets from humans, pre-trained AI and fine-tuned AI. The marginal effects indicate how the likelihood of $Y_i = 1$ changes with a one-unit increase in each explanatory variable. The average predicted probabilities are calculated by applying the estimated logit model to each sample to obtain the probability of choosing Option 1 and then averaging those probabilities. These values align with the selection rates presented in Table 3, confirming that the logit model accurately predicts actual choice tendencies.

<Table 4: Logit Model Estimation: Humans, Pre-trained AI, and Fine-tuned AI (Gain, 3 Attributes)>
<Table 5: Logit Model Estimation: Humans, Pre-trained AI, and Fine-tuned AI (Loss, 3 Attributes)>

We examine the estimation results for humans. For Question 1, all attributes show statistically significant marginal effects at the 1% level. Specifically, being female increases the probability of choosing Option 1 by 4.1%. A 10-year age increase raises the probability by 1.5%, whereas a JPY 1 million increase in household income reduces the probability by 0.5%. For Question 2, statistically significant marginal effects are found only for age and income, but not for gender. Specifically, a 10-year age increase raises the probability of choosing Option 1 by 2.1% and a JPY 1 million increase in income decreases the probability by 0.8%.

Subsequently, we analyze pre-trained AI results. In this setting, statistically significant marginal effects at the 1% level are observed for all three attributes for both questions. For Question 1, being female increases the probability by 3.3%, a 10-year age increase raises the probability by 5.7%, whereas a JPY 1 million increase in income decreases the probability by 2.7%. For Question 2, being female increases the probability by 2.8%, a 10-year age increase raises the probability by 2.3%, whereas a JPY 1 million increase in income decreases the probability by 0.6%.

We then compare the marginal effects of humans and pre-trained AI. Using the delta method, we

calculate the standard errors for the marginal effects and conduct statistical tests to determine the differences between the two groups. For Question 1, the z-values for the differences in marginal effects between humans and pre-trained AI for gender, age, and income are 0.70, -9.90, and 15.56, respectively. The difference in gender is not statistically significant at the 5% level. In contrast, significant differences are observed in terms of age and income. For Question 2, the z-values for the differences in marginal effects for gender, age, and income are -2.04, -0.31, and -0.71, respectively. Unlike in the gain domain, only gender shows a statistically significant difference, while age and income do not show any significant differences. These results suggest that while partial differences in marginal effects between humans and pre-trained AI are observed, no consistent or statistically significant differences are found.

Next, we examine the estimation results for fine-tuned AI. Question 1 confirms statistically significant marginal effects at the 1% level only for age and income. Specifically, a 10-year age increase raises the probability by 3.4%, and a JPY 1 million increase in income decreases the probability by 1.9%. By contrast, statistically significant marginal effects are observed only for gender and age in Question 2. Specifically, being female increases the probability of choosing Option 1 by 9.5%, and a 10-year age increase raises the probability by 2.3%. For Question 2, a marginal effect becomes larger for gender.

Moreover, we compare the responses of humans and those of fine-tuned AI. For Question 1, the z-values for the differences in the marginal effects for gender, age, and income are 3.24, -4.48, and 9.90, respectively; all are statistically significant. For Question 2, the z-values for the differences for gender, age, and income are -5.10, -0.28, and -3.54. Statistically significant differences at the 5% level are found only for gender and income.

Comparing the responses of pre-trained AI and those of fine-tuned AI, for Question 1, the z-values for the differences in the marginal effects for gender, age, and income are -2.93, 5.42, and -5.66, respectively; all are statistically significant. For Question 2, the z-values for the differences for gender, age, and income are -4.03, 0.04, and -2.83. Statistically significant differences at the 5% level are found only for gender and income.

5.3. Graphical Analysis of Choice Probabilities

Figures 1 and 2 illustrate the choice probabilities for each attribute in Questions 1 and 2, respectively. Figure 1 shows that under “Gender,” the other two attributes are fixed at their average values (age = 44.9 years, income = JPY 5.93 million).

<Figure 1: Choice Probabilities: Humans, Pre-trained AI, and Fine-tuned AI (Gain, 3 Attributes)>

<Figure 2: Choice Probabilities: Humans, Pre-trained AI, and Fine-tuned AI (Loss, 3 Attributes)>

In Figure 1, in "Gender," the 95% confidence intervals for humans’ and pre-trained AI’s lines overlap at both points, indicating no significant difference in the choice probabilities. However, fine-tuned AI’s

line is positioned above both, suggesting a stronger risk aversion through training. Meanwhile, in "Age" and "Income," the slope of fine-tuned AI's line becomes more moderate through fine-tuning, approaching that of humans. Nevertheless, near the average values, fine-tuned AI lies above the 95% confidence intervals for both humans and pre-trained AI, again indicating a heightened risk aversion.

Figure 2 shows a different trend. In "Gender" and "Income," the 95% confidence intervals for humans' and fine-tuned AI's lines overlap at minimum value; the slope of fine-tuned AI's line becomes steeper, deviating from that of humans. On the other hand, the line is still closer to that of humans than that of pre-trained AI, indicating that it has effectively learned a human-like risk-seeking behavior in the loss domain. In "Age," no major slope differences are found among the three lines, but apparent differences in y-intercepts are observed. Pre-trained AI's line is above the other two and fine-tuned AI's line is closer to that of humans. In addition, by comparing the scale of the y-coordinates in both figures, we can assess the tendency toward reference dependence. Notably, we demonstrate that training reduces the asymmetric response pattern in both gain and loss domains, resulting in behavior more consistent with human-like reference dependence.

5.4. Analysis of Prediction Accuracy

Table 6 shows the prediction accuracy between the predicted choice probabilities from each model and the actual binary choices. "Predicted value" is defined as a binary classification based on the expected probability $\widehat{\mathrm{P}}(\mathrm{Y_i} = 1|\,\mathbf{X})$. If the predicted probability is 0.5 or greater, the respondent is classified as choosing Option 1; otherwise, they are classified as choosing Option 2. Based on this classification, we calculate the match rate between the model's prediction and the actual observed response to assess the model's fit. In this context, "prediction accuracy" refers to the proportion of cases in which the predicted outcomes from the model match the actual observed outcomes. Specifically, "prediction accuracy" is the sum of the proportions of cases in which both the predicted and observed outcomes are Option 1 (i.e., (1,1)) and those in which both are Option 2 (i.e., (2,2)).

In the human web-based survey results, the prediction accuracy for Question 1 is 88.40%. However, for Question 2, the prediction accuracy decreases to 57.59%. In comparison, the prediction accuracy based on the responses generated by pre-trained AI ranges from 88.98% to 92.08%, indicating a higher overall level of consistency compared with the web-based survey. On the other hand, in the fine-tuned AI results, the prediction accuracy for Question 1 is 94.61%. However, for Question 2, the prediction accuracy decreases to 62.05%, indicating similar tendencies as humans.

<Table 6: Logit Model Prediction Accuracy: Humans, Pre-trained AI, and Fine-tuned AI >

5.5. Comparative Analysis of Choice Probabilities

For the quantitative evaluation of the proximity of decision tendencies between humans and generative AI before and after fine-tuning, we utilize the t-test based on average choice values introduced in Section 5.1. We also report the Brier score[14] and log loss[15] as supplementary measures of predictive performance. Both the Brier score and log loss are standard metrics widely used in machine learning to evaluate predictive accuracy by measuring the discrepancy between observed outcomes and predicted probabilities. In calculating these measures, Y_i denotes the observed human choice (0 or 1) while $\hat{p}_i$ denotes the predicted probabilities estimated by the logit model corresponding to the AI's response. Table 7 presents these three indicators for comparisons between humans and pre-trained AI, as well as between humans and fine-tuned AI.

<Table 7: The Proximity of AI to Humans through Fine-Tuning >

For Question 1, the t-statistics, Brier score, and log loss for the difference between humans and pre-trained AI are 4.41, 0.122, and 0.487, respectively. The corresponding values for the difference between humans and fine-tuned AI are 10.2, 0.115, and 0.496, respectively. These results indicate that in terms of the t-statistic and log loss, the risk preference tendencies of humans and AI become more divergent after fine-tuning, whereas in terms of the Brier score, they become closer. As the difference in tendencies between humans and pre-trained AI is already small in the gain domain, these findings suggest that fine-tuning may increase the divergence in tendencies between humans and fine-tuned AI in the gain domain.

For Question 2, the t-statistic, Brier score, and log loss for the difference between humans and pre-trained AI are −55.7, 0.461, and 1.34, respectively; the corresponding values for the difference between humans and fine-tuned AI are 4.33, 0.248, and 0.690, respectively. These results indicate that in the loss domain, the risk preference tendencies of humans and AI become closer after fine-tuning from all three evaluative perspectives. Given that the divergence in tendencies between humans and pre-trained AI is substantial in the loss domain, these results indicate that fine-tuning reduces the gap in tendencies between humans and the fine-tuned AI.

Finally, using the combined data from Questions 1 and 2 (N = 9,676), the Brier score and log loss for the difference between humans and pre-trained AI are 0.291 and 0.915, respectively, those for the difference between humans and fine-tuned AI are 0.182 and 0.593, respectively. Collectively, these results suggest that although the effects of fine-tuning are not uniform across the gain and loss domains, fine-tuning overall—particularly in the loss domain—moves generative AI in a direction closer to human tendencies in risk preference and reference dependence[16].

---

[14] The Brier score is given by $n^{-1} \sum_{i=1}^{n} (Y_i - \hat{p}_i)^2$.

[15] The log loss is given by $-n^{-1} \sum_{i=1}^{n} (Y_i \, log(\hat{p}_i) + (1 - Y_i) \, log(1 - \hat{p}_i))$.

[16] For the loss domain and pooled data, the Brier score and log loss measuring the difference between

6. Discussions

6.1. Bias Patterns in Humans and Pre-Trained AI

Based on the analysis presented in Section 5, pre-trained AI exhibits a risk-averse preference pattern in the gain domain, closely resembling that of human respondents. Pre-trained AI demonstrates a high level of consistency in replicating human decision-making behavior. In particular, for Question 1, the selection rate for Option 1 is approximately 90% for both humans and pre-trained AI, suggesting that the behavior aligns with the predictions of prospect theory. Meanwhile, a comparison of the marginal effects by individual attributes reveals no substantial differences between humans and pre-trained AI in terms of gender. However, for age and income, the absolute values of the marginal effects are consistently larger for pre-trained AI, indicating stronger sensitivity to these attributes. Furthermore, pre-trained AI exhibits risk-averse behavior in the loss domain, with a more pronounced asymmetry in preference than humans.

Generative AI tends to replicate the association between demographic attributes and preferences in its training data with excessive precision, suggesting a form of "overfitting to bias." Rather than reflecting the variability and inconsistency inherent in human decisions, the model selects contextually plausible responses, which reinforces stereotypical patterns linked to individual characteristics. The model's responses heavily rely on the context and attribute conditions. This reliance creates a fundamental divergence from human behavior, particularly in how AI reproduces biased structures.

Meanwhile, pre-trained AI exhibits a significantly greater degree of reference dependence than humans. In the web-based survey targeting human respondents, 57.7% choose Option 1 in Question 2, indicating a moderate tendency toward risk aversion. This rate exceeds the 36% reported by Tversky and Kahneman (1988), suggesting a general inclination to avoid losses. However, pre-trained AI selects Option 1 in only 11.0%, demonstrating a more pronounced reference dependence.

Moreover, pre-trained AI reveals a similar pattern of reference dependence. When comparing the selection rates for Option 1 in Questions 1 and 2, Tversky and Kahneman (1988) report rates of 72% and 36%, respectively. In contrast, the rates in our web-based survey are 88% and 58%, respectively. In contrast, pre-trained AI shows a more polarized pattern, with approximately 90% of the responses for Question 1 and 10% for Question 2. This suggests a stronger aversion to losses than observed among human respondents. Furthermore, for Question 2, the comparison of prediction accuracy highlights this

---

humans and the pre-trained AI are statistically significantly larger than those measuring the difference between humans and the fine-tuned AI ($p < 0.01$). In the gain domain, the Brier score for the difference between humans and the pre-trained AI is also statistically significantly greater than that for the difference between humans and the fine-tuned AI ($p < 0.01$), whereas the log loss is statistically significantly smaller ($p < 0.01$).

discrepancy (Table 6): 98.04% of the human responses aligned with predicted choice 1. At the same time, pre-trained AI consistently produces predicted choice two across all samples. These results indicate a clear divergence between humans and pre-trained AI in terms of risk preference and reference dependence.

In addition to GPT-4o, we present the same set of questions to the following LLMs: GPT-3.5 by OpenAI, Gemini 2.5 Flash by Google, and DeepSeek-V3 by DeepSeek. The results reveal differences in response patterns across the models. For example, comparisons between GPT-3.5 and GPT-4o reveal differences in choice tendencies in the gain and loss domains. In addition, Gemini and DeepSeek exhibit weaker risk preference tendencies than GPT-4o in both domains. For a summary of the comparative results, see Appendix A.5.

We also present the pre-trained GPT-4o with risk preference questions in which the monetary stakes are increased tenfold, as well as questions concerning time preference. For the risk preference questions, the responses are broadly consistent with the main findings of this study, suggesting that generative AI exhibits stronger risk preference tendencies than humans (see Appendix A.6 for details). For the questions on time preference, the results suggest relatively weak time preference bias for both humans and AI.

6.2. Bias Patterns in Fine-Tuned AI

The study's results indicate that GPT-4o overfits the human tendency toward risk aversion in the gain domain. For Question 1, the proportions of respondents selecting Option 1 are 88%, 91%, and 94% for humans, pre-trained AI, and fine-tuned AI, respectively. These results reflect the risk-averse preferences predicted by prospect theory; however, fine-tuning may intensify this bias. This pattern supports the earlier claim that the model overfits cognitive biases. When comparing the marginal effects by demographic attributes, the gender gap between humans and AI widens after fine-tuning. However, the gaps related to age and income narrow. Nonetheless, statistically significant differences remain between humans and fine-tuned AI for all three attributes. Using t-statistics and Brier score, we cannot see that AI's risk preference tendencies approach those of humans. However, using log loss, we find that fine-tuning reduces the distance between the human and AI choice distributions.

In the loss domain, fine-tuning reduces the extreme risk-seeking of AI and brings its behavior closer to that of humans. Before training, the pre-trained AI selects Option 1 in Question 2 at a rate of only 11.0%, indicating a strong preference for risk proneness. After fine-tuning, this rate increases to 62.1%, exceeding the 36% reported by Tversky and Kahneman (1988) and approaching the 57.7% observed in our web-based survey. Using all three metrics: t-statistics, Brier score, and log loss, we find that fine-tuning reduces the distance between the human and AI choice distributions., suggesting that learning progresses more effectively in the loss domain than in the gain domain. Furthermore, classification accuracy improves after training. For question 2, the model no longer assigns all samples to Prediction

2. Instead, all of the samples receive Prediction 1. This shift indicates a substantial alignment with human behavior. However, a comparison of the marginal effects by attribute reveals a different pattern. It shows that fine-tuning increases the gap between humans and AI. In particular, the fine-tuned AI yields large marginal effects for gender (0.095), suggesting potential overfitting in the model's response to demographic information.

Additionally, fine-tuning reduces the AI's degree of reference dependence, bringing it closer to human behavior. Comparing the choice rates for Option 1 in Questions 1 and 2, pre-trained AI shows a 90% vs. 10% split. After training, the pattern shifts to 94% vs. 62%. By comparison, Tversky and Kahneman (1988) report a 72% vs. 36% split, whereas the human respondents in this study show an 88% vs. 58% split. These results indicate that fine-tuning weakens the AI's excessive reference dependence and aligns its behavior more closely with human responses.

7. Conclusions

This study examines the decision-making tendencies of humans and generative AI (GPT-4o) in terms of risk preferences and reference dependence. It focuses on the impact of persona settings, defined by attribute information, on AI responses. Furthermore, we examine whether generative AI can replicate human behavioral patterns by incorporating human decision data and evaluate the extent to which fine-tuning improves the alignment between AI and human tendencies.

The comparison of humans and pre-trained AI reveals no substantial difference in the gain domain, where both humans and generative AI exhibit risk-averse tendencies. In the loss domain, generative AI displays stronger risk-seeking tendencies than humans. Regarding the effects of individual attributes such as age and income, generative AI tends to exhibit stronger effects on risk aversion than those observed in humans.

Meanwhile, the comparison of humans and the fine-tuned AI shows that fine-tuning induces changes in a direction such that generative AI tendencies move closer to human tendencies. In particular, in the loss domain, a tendency toward a reduction in the gap between humans and AI is observed, suggesting an overall learning effect. At the same time, some limitations remain: in the gain domain, evidence of overshooting in risk aversion is observed; in the loss domain, the effects of certain individual attributes appear to be overstated, indicating room for further improvement.

References

[1] Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In International Conference on Machine Learning (pp. 337-371). PMLR.

[2] Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2025). Playing repeated games with large language models. Nature Human Behaviour, 1-11.

[3] Ambrosio, L., Gigli, N., & Savaré, G. (2008). Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media.

[4] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. Political Analysis, 31(3), 337-351.

[5] Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6), e2218523120.

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

[7] Brand, J., Israeli, A., & Ngwe, D. (2023). Using GPT for market research. Harvard business school marketing unit working paper, (23-062).

[8] Brookins, P., & DeBacker, J. (2024). Playing games with GPT: What can we learn about a large language model from canonical strategic games?. Economics Bulletin, 44(1), 25-37.

[9] Brown, A. L., Imai, T., Vieider, F. M., & Camerer, C. F. (2024). Meta-analysis of empirical estimates of loss aversion. Journal of Economic Literature, 62(2), 485-516.

[10] Bybee, L. (2023). Surveying Generative AI's Economic Expectations. arXiv preprint arXiv:2305.02823.

[11] Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F., Brañas-Garza, P., Butera, L., Douglas, K. M., Everett, J. A. C., Gigerenzer, G., Greenhow, C., Hashimoto, D. A., Holt-Lunstad, J., Jetten, J., Johnson, S., Kunz, W. H., Longoni, C., Lunn, P., Natale, S., Paluch, S., Rahwan, I., Selwyn, N., Singh, V., Suri, S., Sutcliffe, J., Tomlinson, J., van der Linden, S., Van Lange, P. A. M., Wall, F., Van Bavel, J. J., & Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS nexus, 3(6), pgae191.

[12] Capraro, V., Di Paolo, R., & Pizziol, V. (2025). A publicly available benchmark for assessing large Language models' ability to predict how humans balance self-interest and the interest of others. Scientific Reports, 15(1), 21428.

[13] Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of GPT. Proceedings of the National Academy of Sciences, 120(51), e2316205120.

[14] Chen, Y., Kirshner, S. N., Ovchinnikov, A., Andiappan, M., & Jenkin, T. (2025). A manager and

an AI walk into a bar: does ChatGPT make biased decisions like we do?. Manufacturing & Service Operations Management.

[15] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1, pp. 4171-4186.

[16] Demszky, D., Yang, D., Yeager, D.S., Bryan, C.J., Clapper, M., Chandhok, S., Eichstaedt, J.C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D.C., Dweck, C.S., Gross, J.J., & Pennebaker, J.W. (2023). Using large language models in psychology. Nature Reviews Psychology, 2(11), 688-701.

[17] Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2025). Questioning the survey responses of large language models. Advances in Neural Information Processing Systems, 37, 45850-45878.

[18] Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. The quarterly journal of economics, 133(4), 1645-1692.

[19] Guo, F. (2023). GPT in game theory experiments. arXiv preprint arXiv:2305.05516.

[20] Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. arXiv preprint arXiv:2303.13988, 1.

[21] Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science, 3(10), 833-838.

[22] Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? (No. w31122). National Bureau of Economic Research.

[23] Jia, J., Yuan, Z., Pan, J., McNamara, P., & Chen, D. (2024). Decision-making behavior evaluation framework for llms under uncertain context. Advances in Neural Information Processing Systems, 37, 113360-113382.

[24] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 363-391.

[25] Kitadai, A., Lugo, S. D. R., Tsurusaki, Y., Fukasawa, Y., & Nishino, N. (2024). Can AI with high reasoning ability replicate human-like decision making in economic experiments?. arXiv preprint arXiv:2406.11426.

[26] Lee, S., Peng, T. Q., Goldberg, M. H., Rosenthal, S. A., Kotcher, J. E., Maibach, E. W., & Leiserowitz, A. (2024). Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. PLOS Climate, 3(8), e0000429.

[27] Li, P., Castelo, N., Katona, Z., & Sarvary, M. (2024). Frontiers: Determining the validity of large language models for automated perceptual analysis. Marketing Science, 43(2), 254-266.

[28] Lorè, N., & Heydari, B. (2024). Strategic behavior of large language models and the role of game structure versus contextual framing. Scientific Reports, 14(1), 18490.

[29] Macmillan-Scott, O., & Musolesi, M. (2024). (Ir) rationality and cognitive biases in large language models. Royal Society Open Science, 11(6), 240255.

[30] Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. Proceedings of the National Academy of Sciences, 121(9), e2313925121.

[31] Monarch, R. M. (2021). Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. Simon and Schuster.

[32] Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: measuring ChatGPT political bias. Public Choice, 198(1), 3-23.

[33] Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Willer, R., Liang, P., Bernstein, M.S. (2024). Generative agent simulations of 1,000 people. arXiv preprint, arXiv:2411.10109.

[34] Qiu, L., Singh, P. V., & Srinivasan, K. (2023). Consumer Risk Preferences Elicitation From Large Language Models. Available at SSRN 4526072.

[35] Rahwan (2018). Society-in-the-Loop: Programming the Algorithmic Social Contract. Ethics and Information Technology, 20, 5–14.

[36] Rahwan et al. (2019). Machine Behaviour. Nature, 568, 477–486.

[37] Ross, J., Kim, Y., & Lo, A. W. (2024). Llm economicus? mapping the behavioral biases of llms via utility theory. arXiv preprint, arXiv:2408.02784.

[38] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In International Conference on Machine Learning, pp. 29971-30004.

[39] Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. Psychology & Marketing, 41(6), 1254-1270.

[40] Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., & Chen, H. (2025). Database report: Twin-2k-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. Marketing Science, 44(6), 1446-1455.

[41] Tsuchihashi, T. (2023). How much do you bid? Answers from ChatGPT in first-price and second-price auctions. Journal of Digital Life, 3.

[42] Tversky, A., & Kahneman, D. (1988). Rational choice and the framing of decisions. In D. E. Bell, H. Raiffa, & A. Tversky (Eds.). In Decision making: Descriptive, normative, and prescriptive interactions, pp. 167–192.

[43] Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty, 5, 297-323.

[44] Xie, Y., Liu, Y., Ma, Z., Shi, L., Wang, X., Yuan, W., ... & Mei, Q. (2024). How Different AI

Chatbots Behave? Benchmarking Large Language Models in Behavioral Economics Games. arXiv preprint arXiv:2412.12362.

[45] Xiong, X., Wong, I. A., Huang, G. I., & Peng, Y. (2024). Understanding AI-generated experiments in tourism: replications using GPT simulations. Journal of Travel Research, 00472875241275945.

[46] Wada, S. (1996). Construction of the Big Five Scales of personality trait terms and concurrent validity with NPI. Japanese Journal of Psychology.

[47] Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. Nature Machine Intelligence, 7(3), 400-411.

Table 1: Prior Studies on Decision-Making and Cognitive Biases: Generative AI vs. Humans

Study ID	Authors	Targeted Bias / Experimental Task	AI Model	Population / Sample Size	Average Reproducibility	Conditional Reproducibility	Human Feedback	Evaluation Metric
Economic Decision-Making and Behavioral Biases								
1	Akata et al. (2025)	Economic experiments (Prisoner's Dilemma, Battle of the Sexes)	GPT-3, 3.5, 4	LLM only, using varied payoff matrices across 10 rounds each (1,224 trials total)	N/A (LLM-only results). The model performs well in the Prisoner's Dilemma but shows lower performance in the Battle of the Sexes.	N/A	N/A	N/A
2	Chen et al. (2023)	Economic experiments (Risk Preferences, Time Preferences, Social Preferences, Food Choice)	GPT-3.5-Turbo	Humans: 347 participants; LLM: 4 Domain × 25 questions × 100 trials	LLM:CCEI is extremely high at 0.997-0.999, indicating almost all choices satisfy GARP and demonstrate higher rationality than humans (0.96-0.98).	Changes in price presentation (e.g., displaying one unit → displaying multiple units) significantly reduce rationality. Discrete choice format (continuous → presented with choices) increases GARP violations. Strongly dependent on the framing of the problem.	N/A	CCEI（Critical Cost Efficiency Index） Price elasticity of demand
3	Guo (2023)	Economic experiments (Ultimatum Game, Prisoner's Dilemma)	GPT-4 (gpt4-1106-preview)	LLM only (compared with prior studies): Assigned selfish and fairness-oriented personas (UG: 400 trials, PD: 300 trials)	Similar to findings in prior studies	N/A	Prompt adjustment only	N/A
4	Horton (2023)	Economic experiments (Social Preferences: Charness and Rabin, 2002; Fairness: Kahneman et al., 1986; Status Quo Bias: Samuelson and Zeckhauser, 1988; Minimum Wage: Horton, 2023)	GPT-3 (including davinci-003 and others)	Reproduction of results from prior studies	High similarity with existing results	N/A	RLHF applied.	N/A
5	Tsuchihashi (2023)	Economic experiment (bidding behavior in auctions)	GPT-3.5	LLM Only (compared with prior research): 40 rounds each of FPA and SPA	Similar trend to prior research in FPA: Overbidding in FPA, slightly more accurate bidding in SPA	N/A	Prompt adjustment only	N/A
6	Brookins and DeBacker (2024)	Economic experiments (Dictator Game, Prisoner's Dilemma)	GPT-3.5-turbo	LLM Only (compared with prior studies): Dictator Game: 500 rounds, Prisoner's Dilemma: 1,100 rounds	LLMs tend to give fairer and more cooperative responses than humans.	Evaluated based on risk, temptation, and efficiency. Similar tendencies were observed only in the efficiency metric.	Prompt adjustment only	N/A
7	Kitadai et al. (2024)	Economic experiment (Ultimatum Game)	GPT-3.5 (gpt-3.5-turbo-0613), GPT-4 (gpt-4-1106-preview)	LLM only: For each configuration (prompt type, temperature value, and GPT model version), 1000 agents are generated.	Proposer: Adjusting temperature partially replicates human data; GPT improvements move results closer to the theoretical predictions. Responder: Results approach theory but do not fully match experimental data.	N/A	N/A	N/A
8	Jia et al. (2024)	Economic experiments (loss aversion, probability weighting, risk preference)	ChatGPT-4.0-Turbo, Claude-3-Opus, Gemini-1.0-pro	LLM Only (compared with prior studies)	Somewhat similar tendencies to human choices	Bias varies by attributes (e.g., gender); no direct comparison with human responses.	Prompt adjustment only	N/A

9	Lorè and Heydari (2024)	Economic experiments (strategic behavior in two-player games)	GPT-3.5, GPT-4, LLaMa-2	LLM Only: 4 Games × 5 Contexts × 3 Models × 300 Trials	N/A (LLM results only) GPT-4 shows structure-dependent behavior; GPT-3.5 shows context-dependent behavior	N/A	Prompt adjustment only	N/A
10	Mei et al. (2024)	Economic experiments (altruism, fairness, trust, cooperation)	GPT-3.5-Turbo, GPT-4	Compared with public data from approximately 90,000 human decisions, LLM data: 6 games × 30 rounds.	ChatGPT-4 produces results similar to those of humans.	N/A	Prompt adjustment only	N/A
11	Qiu et al. (2024)	Insurance plan selection	GPT-4-turbo	Western insurance data up to 2006 (5,998 cases)	At the aggregate level, the model exhibits choice patterns similar to those of humans; however, it performs poorly at the individual level.	Psychological parameters estimated from LLM responses, such as loss aversion and the probability weighting coefficient, tend to have smaller values compared to those reported in previous human studies.	N/A	F1score
12	Ross et al. (2024)	Economic experiments (inequity aversion, loss aversion, and time discounting)	GPT-3.5, GPT-4, Claude 2, and Nine Other LLMs	LLM Only (compared with prior studies)	Behavioral differences from Humans: Inequity aversion: LLMs exhibit intense guilt toward others but weak envy. Loss aversion: they respond rationally to gains but show irrational tendencies toward losses. Temporal discounting: they display a more substantial present bias than humans.	N/A	Prompt adjustment only	N/A
13	Xie et al. (2024)	Economic experiments (altruism, fairness, risk, cooperation)	GPT-4o, LLaMa3, Claude 3, and others	LLMs only (compared with prior studies): 6 games × 5 models × 50 trials	The distribution of LLM responses showed patterns similar to those of humans. The models tended to make fair choices and exhibited high rates of cooperation.	N/A	Prompt adjustment only	Wasserstein distance
14	Capraro et al. (2025)	Economic experiment (Dictator Game)	GPT-4, GPT-4o, Bard, Bing chat	LLM Only (compared with human data (average and distribution) from 38 prior studies in 12 countries)	Tendency to overestimate average donation amount (approximately 30% in actual data → over 40% in GPT). Other methods besides GPT perform poorly.	The distribution shape (three peaks at 0%, 50%, and 100%) is accurately reproduced. 100% overestimates altruistic behavior, 0% underestimates selfish behavior. Responds to framing, but the level is always high.	N/A	Mean error Distribution error
15	Chen et al. (2025)	18 types of cognitive biases	GPT-3.5-turbo, GPT-4	LLM only (binary classification of presence/absence of bias)	Many biases are replicated. GPT-3.5 tends to avoid losses but prefers risk in gains. GPT-4 exhibits consistent risk aversion across various framing contexts.	N/A	N/A	N/A
16	Macmillan-Scott and Musolesi (2024)	12 cognitive tasks	GPT-3.5, GPT-4, Claude2, Bard, LLaMA	LLM Only (compared with prior studies)	GPT-3.5 exhibits the highest proportion of human-like biases. GPT-4 shows the most human-like response patterns overall.	N/A	N/A	N/A
Reproducibility of Survey Data								

17	Santurkar et al. (2023)	Public opinion survey data	GPT models, AI21 Labs models	United States (ATP public opinion surveys, 2017–2021)	The opinions generated by language models exhibit significant discrepancies compared to those of the general U.S. population. They tend to exhibit a specific political bias, leaning toward liberal viewpoints.	N/A	RLHF applied.	Wasserstein distance
18	Park et al. (2024)	Personality assessments and behavioral experiments (15 types of economic and psychological experiments)	Agent architectures utilizing LLMs	Simulation of 1,052 AI agents	Reproduced participants' responses two weeks later with high accuracy (approximately 85%).	Evaluated under conditions based on age, race, and political ideology. Showed a consistent tendency to reduce bias across tasks.	Prompt adjustment only	N/A
19	Dominguez-Olmedo et al. (2025)	Reproduction of survey data (order and label bias)	GPT-2 to GPT-4, LLaMA, and 43 other models	25 questions from the 2019 American Community Survey (ACS)	Overall, the accuracy is low.	N/A	RLHF applied.	KL divergence
20	Toubia et al. (2025)	Behavioral economics experiments + Psychology, Cognition, and Preferences	GPT-4.1, GPT-4.1-mini, Gemini-flash2.5	4-wave panel survey (2,058 people) Wave 1-3: Basic survey Wave 4: Follow-up survey	Individual-level prediction accuracy: Approximately 71.7% Achieves approximately 88% of the accuracy of human re-examination (approximately 81.7%) Average behavior can be reproduced with high accuracy.	There are cases where irrational behavior cannot be reproduced. There is a disconnect from humans in the medical and political fields (e.g., vaccine refusal: 45% of humans vs. 4% of LLMs). There is an overreaction to correct answers in knowledge-based questions.	N/A	Prediction accuracy and Reproduction of ATE (Average Treatment Effect)
Cognitive and Psychological Biases								
21	Aher et al. (2023)	Turing experiments (Ultimatum Game, Garden Path Sentences, Milgram Shock Experiment, and Wisdom of Crowds)	GPT-3.5, GPT-4, and 8 other models	LLMs only (compared with previous studies): 1,000 virtual subjects per task, several thousand responses in total.	Models from LM-5 onward show a higher level of resemblance to human responses. However, instances of hyper-accuracy bias are observed, particularly in the wisdom of crowds task.	N/A	RLHF applied.	N/A
22	Binz and Schulz (2023)	Psychological experiments (decision-making ability, information search ability, deliberative capacity, and causal reasoning ability)	GPT-3 (Davinci)	LLMs only (compared with prior studies): repeated testing with thousands of questions (13,000 responses).	Some responses, including incorrect answer patterns, exhibit characteristics similar to those of humans.	N/A	Prompt adjustment only	N/A
23	Hagendorff et al. (2023)	Psychological experiments (CRT, semantic illusion)	GPT-1 to GPT-4	Humans: 455 participants; LLMs: 50 trials per task across 10 model types	Up to GPT-3, responses tend to be intuitive; GPT-4 has a higher accuracy than humans, achieving a correct response rate of 96%.	N/A	Prompt adjustment only	N/A

24	Wang et al. (2025)	Social Identity, Stereotypes, and Value Judgments (16 attributes including race, gender, disability, and politics)	GPT-4 etc.	Humans: 3,200 participants; LLMs: Answer by fully embodying a specific attribute.	Overall, the reproducibility is low. In particular, it fails to reproduce the perspective of those involved and is biased towards an external perspective.	Misportrayal: Reproduces external stereotypes rather than those of the individuals involved. Flattening: Fails to reproduce diversity within the group, resulting in homogenized responses. Ssentialization: Simplifies attributes into fixed ones. Distortion is particularly pronounced in minority groups.	N/A	Jackard distance, Cosine distance, Wasserstein distance, and Difference of average values
Political Attitudes and Voting Behavior								
25	Argyle et al. (2023)	Political attitudes and voting behavior	GPT-3	ANES: 1,304 individuals + 2,873 individuals	A strong correlation (greater than 0.9) is observed in voting behavior.	Reproducibility is low in specific categories, such as independents.	Prompt adjustment only	Cramér's V, Tetrachoric Correlation
26	Bisbee et al. (2024)	Comparison of ANES data using political persona settings	ChatGPT 3.5 Turbo, ChatGPT 4.0, Falcon-40B	United States (ANES 2016 and 2020)	LLM responses closely resemble human responses but exhibit less variability.	When examined by political attributes, human and LLM responses show similar overall tendencies despite differing variability. However, when conditioned on individual attributes, human and LLM responses tend to diverge in their patterns.	RLHF applied.	N/A
27	Motoki et al. (2024)	Political bias	GPT-3.5	LLM only: 100 responses under Democratic/Republican persona prompts and 100 responses without persona prompts	No comparison with humans. Default responses tend to align with Democratic positions. The political stance changes depending on the specified persona.	N/A	Prompt adjustment only	N/A
Other Studies								
28	Brand et al. (2023)	Willingness to pay (WTP) for multiple products	GPT-3.5-turbo-0125	LLM only (compared with prior studies)	The average WTP patterns are well replicated. However, for new products, consistency with human responses tends to decline.	For specific product categories, the model exhibits patterns similar to those of humans.	Fine-tuning applied.	N/A
29	Bybee (2023)	Economic expectations forecast (based on WSJ articles)	GPT-3.5	United States (300 WSJ articles from 1984 to 2021)	High correlation with existing surveys such as SPF, AAII, and CFO.	N/A	N/A	N/A
30	Lee et al. (2024)	Beliefs about global warming	GPT-3.5-turbo-16k, GPT-4	U.S. national surveys (2017: 1,304 respondents; 2021: 1,006 respondents)	Beliefs about global warming are highly replicable (85%), while replication rates for beliefs about its causes and associated concerns are lower (51% and 48%, respectively).	The opinions of Black respondents tend to be underrepresented.	Prompt adjustment only	F1 score, Cramér's V
31	Li et al. (2024)	Brand recognition and perceived similarity.	GPT-4, GPT-Neo	Evaluations of 21 automobile brands by 530 participants.	High accuracy rate (87.2%).	Consistent patterns across age groups and demographic attributes.	Prompt adjustment only	Triplet matching rate
32	Xiong et al. (2024)	Emotion and belief formation (in the tourism domain).	GPT-3.5-turbo	LLM only (compared with prior research): 16 scenarios × 100 iterations.	The responses generated by the LLM exhibit trends similar to those observed in human responses.	N/A	Prompt adjustment only	N/A

Table 2: Descriptive Statistics

Individual Attributes	Mean
Female dummy	0.507 [0.500]
Age (10 years)	4.49 [1.33]
Annual income (JPY 1 million)	5.93 [3.10]
Obs.	4,838

Notes: The values in parentheses in the table indicate standard deviations.

Table 3: Choice Results: Humans, Pre-trained AI, and Fine-tuned AI

Question 1 Gain	Pre-trained AI temperature 1.0	Fine-tuned AI temperature 1.0	Human
Mean	0.911	0.942	0.884
Std. Dev.	0.285	0.234	0.320
Obs.	4,838	4,838	4,838

Question 2 Loss	Pre-trained AI temperature 1.0	Fine-tuned AI temperature 1.0	Human
Mean	0.110	0.621	0.577
Std. Dev.	0.313	0.485	0.494
Obs.	4,838	4,838	4,838

Table 4: Logit Model Estimation: Humans, Pre-trained AI, and Fine-tuned AI (Gain, 3 Attributes)

Question 1 Gain		Pre-trained AI temperature 1.0	Fine-tuned AI temperature 1.0	Human
Marginal Effects	Female dummy	0.033*** (0.007)	0.006 (0.006)	0.041*** (0.009)
	Age (10 years)	0.057*** (0.003)	0.034*** (0.003)	0.015*** (0.003)
	Annual income (1 million yen)	-0.027*** (0.001)	-0.019*** (0.001)	-0.005*** (0.001)
obs.		4,838	4,838	4,838
McFadden $\mathbf{R^2}$		0.3472	0.3210	0.0140
Predicted Probability		0.911 (0.004)	0.942 (0.003)	0.884 (0.005)

Notes: Rows 2-4 show the marginal effects of each attribute based on the logit model. The numbers in parentheses indicate standard errors. " *** " denotes statistical significance at the 1% level. Row 7 presents the average predicted probabilities based on the estimates from the logit model.

Table 5: Logit Model Estimation: Humans, Pre-trained AI, and Fine-tuned AI (Loss, 3 Attributes)

Question 2 Loss		Pre-trained AI temperature 1.0	Fine-tuned AI temperature 1.0	Human
Marginal Effects	Female dummy	0.028*** (0.009)	0.095*** (0.014)	-0.006 (0.014)
	Age (10 years)	0.023*** (0.004)	0.023*** (0.005)	0.021*** (0.005)
	Annual income (1 million yen)	-0.006*** (0.002)	0.002 (0.002)	-0.008*** (0.002)
obs.		4,838	4,838	4,838
McFadden $\mathbf{R^2}$		0.0204	0.0102	0.0038
Predicted Probability		0.110 (0.004)	0.621 (0.007)	0.577 (0.007)

Notes: Rows 2 to 4 show the marginal effects of each attribute based on the logit model. The numbers in parentheses indicate standard errors. " *** " denotes statistical significance at the 1% level. Row 7 presents the average predicted probabilities based on the estimates from the logit model.

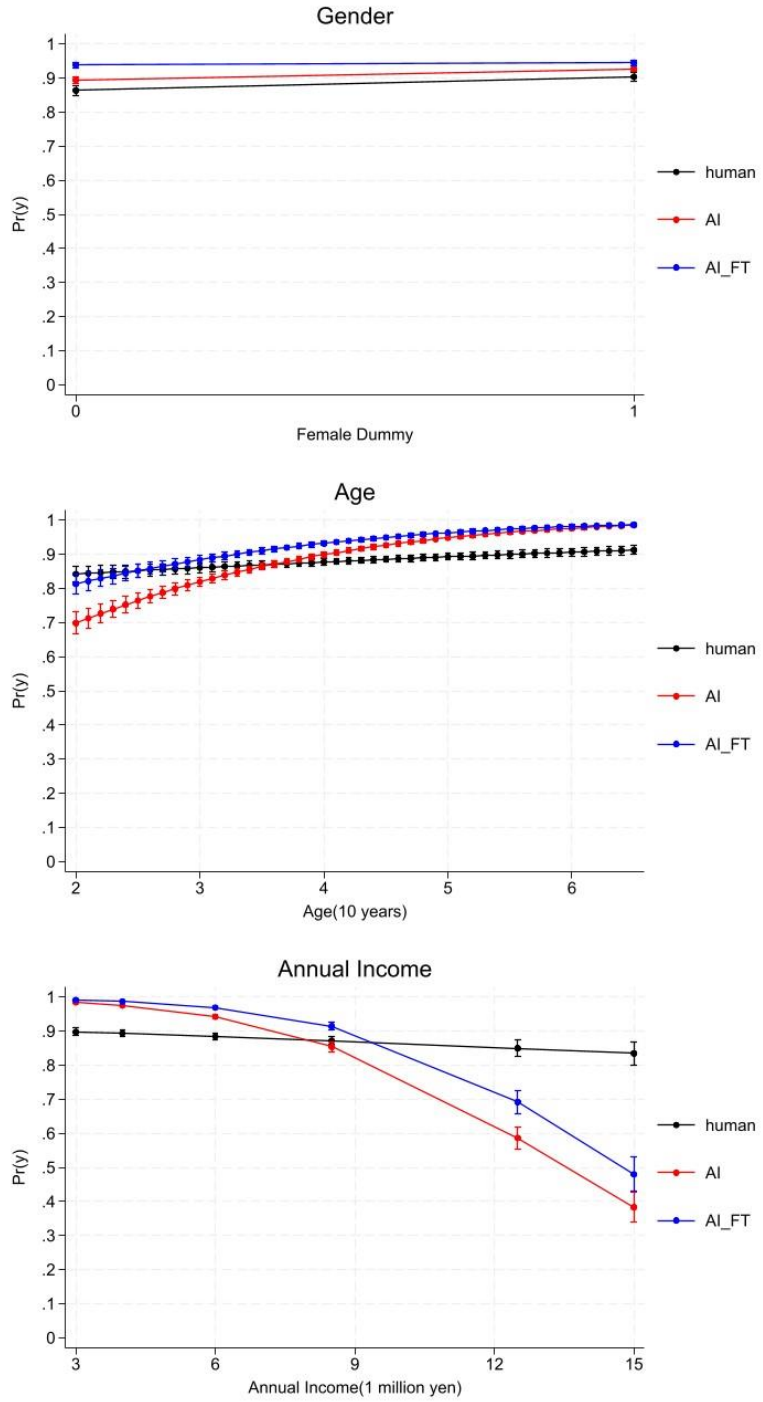

Figure 1: Choice Probabilities: Humans, Pre-trained AI, and Fine-tuned AI (Gain, 3 Attributes)

Notes: AI_10 and AI_FT_10 refer to the responses from the pre-trained AI and the fine-tuned AI, respectively, when asked with a temperature setting of 1.0. The vertical bars at each point in the figure represent 95% confidence intervals.

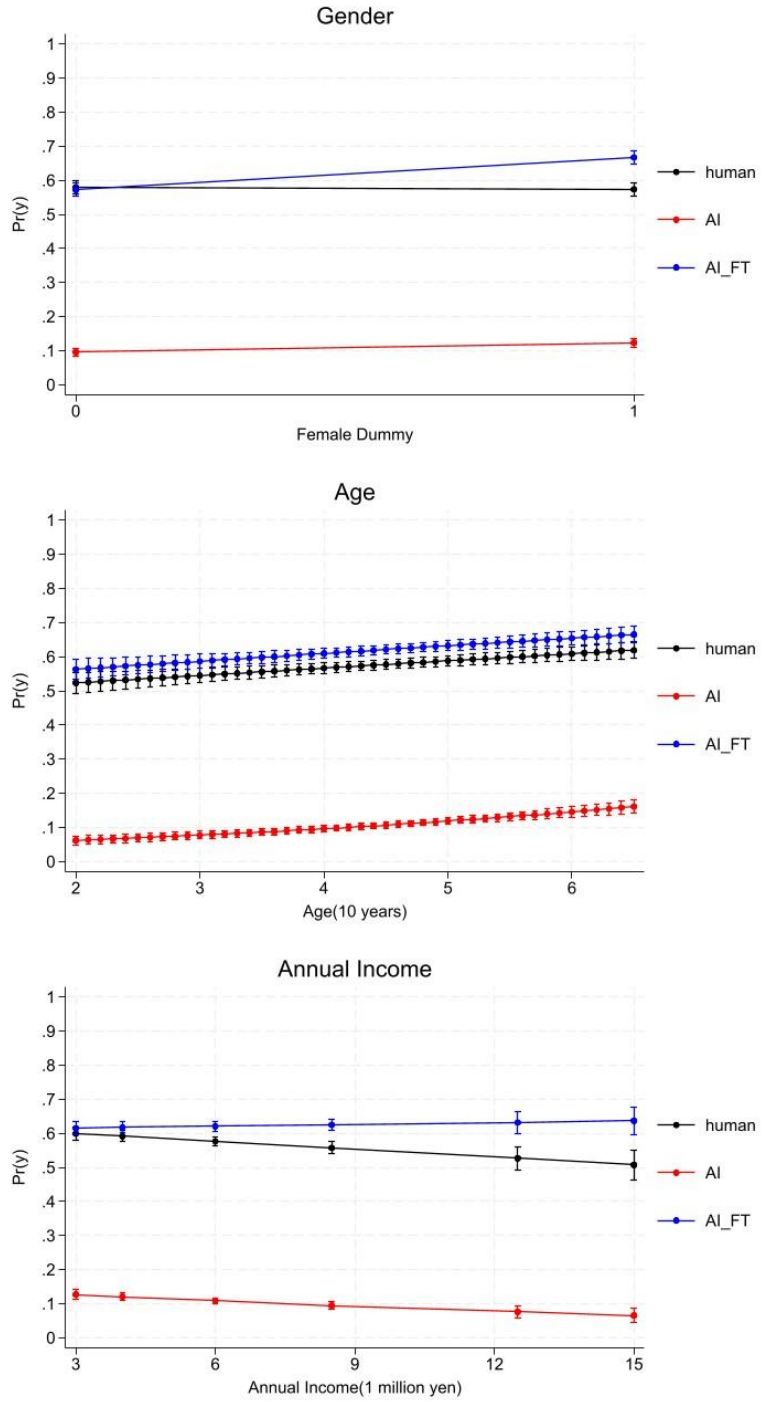

Figure 2: Choice Probabilities: Humans, Pre-trained AI, and Fine-tuned AI (Loss, 3 Attributes)

Notes: AI_10 and AI_FT_10 refer to the responses from the pre-trained AI and the fine-tuned AI, respectively, when asked with a temperature setting of 1.0. The vertical bars at each point in the figure represent 95% confidence intervals.

Table 6: Logit Model Prediction Accuracy:

Humans, Pre-trained AI, and Fine-tuned AI

Question 1 Gain	Pre-trained AI temperature 1.0	Fine-tuned AI temperature 1.0	Human
Predicted value, Observed value			
(2, 2)	2.34% (n=113)	1.32% (n=64)	0% (n=0)
(2, 1)	1.34% (n=65)	0.89% (n=43)	0% (n=0)
(1, 2)	6.57% (n=318)	4.51% (n=218)	11.60% (n=561)
(1,1)	89.75% (n=4,342)	93.28% (n=4,513)	88.40% (n=4,277)
Prediction Accuracy	92.08%	94.61%	88.40%

Question 2 Loss	Pre-trained AI temperature 1.0	Fine-tuned AI temperature 1.0	Human
Predicted value, Observed value			
(2, 2)	88.98% (n=4,305)	0% (n=0)	0.93% (n=45)
(2, 1)	11.02% (n=533)	0% (n=0)	1.03% (n=50)
(1, 2)	0% (n=0)	37.95% (n=1,836)	41.38% (n=2,002)
(1,1)	0% (n=0)	62.05% (n=3,002)	56.66% (n=2,741)
Prediction Accuracy	88.98%	62.05%	57.59%

Notes: The percentage values are rounded to the third decimal place; therefore, the sum of the values in cells (2,2) and (1,1) does not necessarily match the prediction accuracy.

Table 7: The Proximity of AI to Humans through Fine-Tuning

t-value	Between Human and Pre-trained AI	Between Human and Fine-tuned AI
Gain (N=4,838)	4.41	10.2
Loss (N=4,838)	-55.7	4.33

Brier score	Between Human and Pre-trained AI	Between Human and Fine-tuned AI
Gain (N=4,838)	0.122	0.115
Loss (N=4,838)	0.461	0.248
Total (N=9,676)	0.291	0.182

log loss	Between Human and Pre-trained AI	Between Human and Fine-tuned AI
Gain (N=4,838)	0.487	0.496
Loss (N=4,838)	1.34	0.690
Total (N=9,676)	0.915	0.593

Notes: Regarding the t-value, $p < 0.01$ is observed in all cases.

A. Online Appendix

A.1 Analysis with Attribute Information Added

Table A.1.1: Descriptive Statistics (10 Attributes)

Individual Attributes		Mean
Female dummy		0.439 [0.496]
Age (10 years)		4.40 [1.33]
Annual income (JPY 1 million)		5.93 [3.11]
Job		1.53 [0.763]
Education years		14.2 [2.18]
Big Five personality traits	Extraversion	2.91 [0.769]
	Neuroticism	3.30 [0.877]
	Conscientiousness	3.19 [0.670]
	Agreeableness	3.27 [0.614]
	Openness	2.85 [0.730]
Obs.		4,137[17]

Notes: The values in parentheses in the table indicate standard deviations. “Job” is a categorical variable, with full-time workers assigned a value of 1, part-time workers 2, and unemployed/students 3. For “Big Five personality traits,” following Wada (1996), 12 questions are asked for each personality factor, totaling 60 questions, using a 5-point scale from "1. Strongly disagree" to "5. Strongly agree," and the average value is taken.

[17] Respondents who do not answer job, education, or Big Five questions are excluded from the main analysis (n = 701). Therefore, the final sample includes 4,137 participants.

Table A.1.2: Choice Results: Humans vs. Pre-trained AI (10 Attributes)

	Question 1 Gain		Question 2 Loss	
	Pre-trained AI temperature 1.0	Human	Pre-trained AI temperature 1.0	Human
Mean	0.933	0.881	0.364	0.571
Std. Dev.	0.250	0.324	0.481	0.495
Obs.	4,137	4,137	4,137	4,137

Table A.1.3: Logit Model Estimation: Humans vs. Pre-trained AI (10 Attributes)

			Question 1 Gain		Question 2 Loss	
			AI	Human	AI	Human
Marginal Effects	Female dummy		0.022***	0.034***	0.054***	-0.032*
			(0.007)	(0.011)	(0.015)	(0.017)
	Age (10 years)		0.018***	0.016***	0.008	0.014**
			(0.003)	(0.004)	(0.005)	(0.006)
	Annual income (JPY 1 million)		-0.004***	-0.003*	0.002	-0.006**
			(0.001)	(0.002)	(0.002)	(0.003)
	Job (Reference: Full-time)	Part-time	-0.000	-0.004	0.005	0.031
			(0.010)	(0.015)	(0.019)	(0.021)
		Unemployed or Student	-0.000	0.041***	0.012	0.027
			(0.010)	(0.013)	(0.020)	(0.022)
	Education years		0.001	-0.003	0.003	-0.009**
			(0.002)	(0.002)	(0.003)	(0.004)
	Big Five personality traits	Extraversion	-0.097***	0.009	-0.190***	0.018
			(0.007)	(0.009)	(0.012)	(0.013)
		Neuroticism	0.017***	0.013*	0.067***	-0.006
			(0.005)	(0.007)	(0.009)	(0.010)
		Conscientiousness	0.052***	0.029***	0.132***	0.034**
			(0.006)	(0.009)	(0.013)	(0.014)
		Agreeableness	0.044***	0.003	0.087***	-0.025
			(0.007)	(0.010)	(0.014)	(0.015)
		Openness	-0.054***	-0.027***	-0.064***	-0.013
			(0.007)	(0.008)	(0.011)	(0.012)
obs.			4,137	4,137	4,137	4,137
McFadden R^2			0.3591	0.0243	0.1194	0.0067
Predicted Probability			0.933	0.881	0.364	0.571
			(0.002)	(0.001)	(0.003)	(0.001)

Notes: Rows 3-13 show the marginal effects of each attribute based on the logit model. The numbers in parentheses indicate standard errors. " *** ", " ** ", and " * ", denotes statistical significance at the 1%, 5%, and 10 % level, respectively. Row 7 presents the average predicted probabilities based on the estimates from the logit model. Here, all instances of "AI" in the table refer to pre-trained AI. when queried with a temperature setting of 1.0.

A.2 The Process of Implementing “Persona” into AI and Generating AI Responses

First, we extract the three attributes (gender, age, and household income) from web-based survey data and save them in CSV format (Table A.2.1). Next, we explain the process of collecting responses from generative AI. We use Python scripts to generate AI responses simultaneously and efficiently (Figure A.2). Lines 1–5 of the script import necessary libraries (e.g., pandas, csv, and openai) for reading CSV files and communicating with the OpenAI API. Line 8 sets the API key (authentication information) to enable the use of the OpenAI API. Lines 11–12 specify the path to the CSV file (Table A.2.1) containing the analysis targets. Lines 14–31 define the function “extract_reason_and_answer” to extract "reason" and "answer" from the generated responses, outputting "unknown" if no reasoning is provided. Lines 33–74 define the function “generate_reason_and_answer”, which constructs prompts based on each persona’s attributes (gender, age, household income) and generates AI responses via the API. Line 76 applies this process to all 4,838 samples on the DataFrame, storing the extracted reasons and answers as separate variables. The subsequent code saves the response data as CSV files (Table A.2.2) for further analysis.

Table A.2.1: CSV File for Implementing “Persona” into AI

gender	age	income
male	26	400
female	39	850
male	38	400
male	22	850
male	61	400
female	41	850
male	47	300
male	38	850
male	55	300

```python
import pandas as pd
import os
import time
import re
import openai

# "***" denotes API key.
openai.api key = "***"

# "*****" denotes the file name of CSV file containing gender, age, and income information (e.g., Table A.2.1).
file path = "*****.csv"
df = pd.read csv(file path)

def extract reason and answer(response text):
    """
    The function which extracts “reason” and “answer” from OpenAI’s response
    """
    match reason = re.search(r"reason[::]¥s*(.*)", response text)
    match answer = re.search(r"(answer[::]¥s*[12])", response text)

# Even if the format "Reason:" is not present, the first line will be accepted as the reason.
    if not match reason:
        lines = response text.split("¥n")
        reason = lines[0].strip() if lines else "不明"
    else:
        reason = match reason.group(1).strip()

# Even without the format "Answer:", you can get the first "1" or "2".
    answer = match answer.group(1).strip()

    return reason, answer

def generate reason and answer(row):
    """
    The function which generates “reason” and “answer” from AI
    """
     question = f""" The following question is asked to someone with the following gender: {row['gender']}, age: {row['age']}years, annual income: JPY {row['income']} × 10,000.
    Imagine you receive an additional JPY 30,000 on top of your current wealth and are asked to choose between the options below. Which option would you choose?
   Option 1: Receive a guaranteed JPY 10,000
   Option 2: A 50% chance of receiving 20,000 yen and a 50% chance of receiving nothing

    Please write your answer in the following exact format:
    - Reason: [Reason for your choice]
    - Answer: [Output only 1 or 2]
    """

    for attempt in range(3):
        try:
            time.sleep(1)
            response = openai.chat.completions.create(
# "********" denotes the name of GPT model(e.g., gpt-4o-2024-08-06).
                model="********",
                messages=[
                    {"role": "system", "content": f" Your attributes are gender: {row['gender']}, age: {row['age']}years, annual income JPY {row['income']} × 10,000." "},
                    {"role": "user", "content": question}
                ],
                temperature=1.0,
                max tokens=180
            )

            full response = response.choices[0].message.content.strip()
            print(f"✅ API Response for row {row.name}: {full response}")

#For personas where the reason and answer cannot be obtained due to a system issue, display "API Error" in the CSV file.
            if not full response.strip():
                return pd.Series(["API エラー", "API エラー"])

            reason, answer = extract reason and answer(full response)
            return pd.Series([reason, answer])

# For personas where the reason and answer cannot be obtained due to a system issue, display "⚠ API request failed: {e}" in the notebook.
        except Exception as e:
            print(f"⚠ API request failed: {e}")
            return pd.Series(["API Error ", " API Error "])

df[["reason", "answer"]] = df.apply(generate reason and answer, axis=1)

# "**********" denotes the file name of CSV file containing gender, age, income, reason, and answer (e.g., Table A.2.2).
output file path = "**********.csv"
df.to csv(output file path, index=False)

print(f"✅ Generated the reasons and answers and saved them in a new file: {output file path}")
```

Figure A.2: Python Script for Implementing “Persona” into AI and Generating Response

Table A.2.2: CSV File Containing Responses from AI

Gender	Age	Income	Reason	Answer
Male	26	400	Given my current income and age, I feel more secure in receiving a guaranteed gain than taking a risk. Receiving the guaranteed JPY 10,000 gives me peace of mind.	Answer: 1
Female	39	850	By receiving the guaranteed JPY 10,000, I can safely and steadily increase my funds without any risk.	Answer: 1
Male	38	400	Choosing the guaranteed JPY 10,000 is a less risky option, and with a limited income, I prefer stable and secure gains.	Answer: 1
Male	22	850	Considering my current income and age, I am willing to take some risks and make decisions based on expected value. Since the expected value of Option 2 is JPY 10,000, I choose to take the risk.	Answer: 2
Male	61	400	Receiving JPY 10,000 for sure allows me to earn some extra income while avoiding risk, which is reassuring.	Answer: 1
Female	41	850	I value stability, so I prefer an option with a guaranteed return. I chose the option to receive the guaranteed JPY 10,000 because it involves no risk.	Answer: 1
Male	47	300	Given my current income, it feels safer and more beneficial to increase my money in a guaranteed way. I want to prioritize a stable income, so I chose the option that ensures a monthly income of JPY 10,000.	Answer: 1
Male	38	850	Considering my current income and financial situation, I prefer to increase my earnings without taking major risks. Receiving the guaranteed JPY 10,000 helps me maintain financial stability in my life.	Answer: 1
Male	55	300	Since my income is limited, I want the immediate sense of security that comes from receiving the guaranteed JPY 10,000. Avoiding risk and making a safe choice feels more appropriate for me.	Answer: 1

Notes: Table A.2.2 shows excerpts of responses obtained through the procedure described in this Section.

A.3 Comparison of Results due to Temperature Fluctuations

We compare how the response results of pre-trained AI change depending on the setting of the "temperature" parameter, which controls the randomness. First, we examine how the temperature setting affects the response tendencies of pre-trained AI (Table A.3.1). For Question 1, selection rates for Option 1 remained high across different temperatures (91.1% (1.0), 88.9% (0.5), and 89.8% (0.0)), confirming risk aversion in the gain domain. However, for Question 2, selection rates for Option 1 decrease as the temperatures decrease (11.0% (1.0), 3.0% (0.5), and 0.0% (0.0)), suggesting a stronger risk-seeking tendency at lower temperatures. Notably, at temperature = 0.0, only one respondent (persona is "female, age: 53, income: JPY 3 million") selects Option 1. As this response deviates significantly from the broader trend, the data is excluded from subsequent analyses.

Next, we examine how changes in the temperature setting affect the marginal effects of pre-trained AI (Table A.3.2 and A.3.3). In all temperature conditions, statistically significant marginal effects at the 1% level are observed for all attributes for both questions. For Question 1, the marginal effect of age is consistently larger for pre-trained AI than for humans and increases with lower temperatures (0.057 (temp=1.0), 0.073 (temp=0.5), and 0.120 (temp=0.0)) compared with the result for humans (0.015). For income, while the negative marginal effects of AI are larger than those of humans (-0.005), a consistent trend is not found across the different temperature settings (-0.027 (temp = 1.0), -0.033 (temp = 0.5), and -0.031 (temp = 0.0)). For gender, no consistent trend is observed for temperature (Human: 0.041; AI: 0.033 (temp = 1.0), 0.026 (temp = 0.5), and 0.086 (temp = 0.0)). For Question 2, no clear temperature-dependent trend is found in the marginal effect of age (Human: 0.021; AI: 0.023 (temp = 1.0), 0.008 (temp = 0.5)). For income, AI shows slightly stronger negative marginal effects than the human result (-0.008); however, no consistent trend across temperature settings is found as well (both temps: -0.006). For gender, no variation is observed across temperature settings (both temps: 0.028).

Figures A.3.1 and A.3.2 illustrate the choice probabilities for each attribute in Questions 1 and 2, respectively. Figure A.3.1 shows that under "Gender," the other two attributes are fixed at their average values (age = 44.9 years, income = JPY 5.93 million). The choice probabilities are plotted for the gender dummy variable with values of 0 (male) and 1 (female). Additionally, in "Gender," we observe that the 95% confidence intervals overlap at both points. This indicates no significant difference in choice probabilities between the web-based survey results and AI responses, or among the different temperature conditions.

Meanwhile, in "Age" and "Income," the slope of the black line representing the human web-based survey results is relatively gentle, whereas the lines for the generative AI are steeper. This suggests that AI emphasizes changes in the choice probability more strongly in response to age or income. However, the confidence intervals overlap around the mean values in these graphs, indicating that the differences between humans and AI are relatively small near the average. Figure A.3.2 illustrates a clear difference in the y-intercept across all three attributes. The y-intercepts for human web-based

survey results lie around 0.5–0.6, whereas those for the AI results are clustered around 0–0.1. This indicates notable differences in choice probabilities at lower age and income levels. However, the slopes of the lines do not differ significantly between humans and AI, suggesting that incremental changes in the choice probability with increasing age or other attributes are similar.

In addition, by comparing the scale of the y-coordinates in both figures, we can assess the tendency toward reference dependence. In Figure A.3.1, under "Age," the AI's choice probability at age 20 (with temperature = 1.0) is around 0.7. In contrast, in Figure A.3.2, under "Age," the choice probability under the same condition is only about 0.05. This sharp contrast reflects the previously observed strong tendency towards reference dependence in generative AI. Moreover, the figures illustrate that while humans show relatively weaker reference dependence, AI displays this behavior more prominently.

Lastly, we explain the prediction accuracy between the predicted choice probabilities from each model and the actual binary choices (Table A.3.4). In the human web-based survey results, the prediction accuracy for Question 1 is 88.40%. However, for Question 2, the prediction accuracy decreases to 57.59%. In comparison, the prediction accuracy based on the responses generated by the AI ranges from 88.98% to 96.96%, indicating a higher overall level of consistency compared with the web-based survey.

Table A.3.1：Choice Results: Humans vs. Pre-trained AI

Question 1 Gain	AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	Human
Mean	0.898	0.889	0.911	0.884
Std. Dev.	0.302	0.314	0.285	0.320
Obs.	4,838	4,838	4,838	4,838

Question 2 Loss	AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	Human
Mean	0.000	0.030	0.110	0.577
Std. Dev.	0.014	0.172	0.313	0.494
Obs.	4,838	4,838	4,838	4,838

Notes： All instances of "AI" in the table refer to the "pre-trained AI."

Table A.3.2: Logit Model Estimation: Humans vs. Pre-trained AI (Gain, 3 Attributes)

Question 1 Gain		AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	Human
Marginal Effects	Female dummy	0.086***	0.026***	0.033***	0.041***
		(0.006)	(0.007)	(0.007)	(0.009)
	Age (10 years)	0.120***	0.073***	0.057***	0.015***
		(0.004)	(0.003)	(0.003)	(0.003)
	Annual income (1 million yen)	-0.031***	-0.033***	-0.027***	-0.005***
		(0.001)	(0.001)	(0.001)	(0.001)
obs.		4,838	4,838	4,838	4,838
McFadden $\mathbf{R^2}$		0.6355	0.3932	0.3472	0.0140
Predicted Probability		0.898	0.889	0.911	0.884
		(0.004)	(0.005)	(0.004)	(0.005)

Notes: Rows 2-4 show the marginal effects of each attribute based on the logit model. The numbers in parentheses indicate standard errors. " *** " denotes statistical significance at the 1% level. Row 7 presents the average predicted probabilities based on the estimates from the logit model. Here, all instances of "AI" in the table refer to "pre-trained AI."

Table A.3.3: Logit Model Estimation Humans vs. Pre-trained AI (Loss, 3 Attributes)

Question 2 Loss		AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	Human
Marginal Effects	Female dummy	-	0.028*** (0.006)	0.028*** (0.009)	-0.006 (0.014)
	Age (10 years)	-	0.008*** (0.002)	0.023*** (0.004)	0.021*** (0.005)
	Annual income (1 million yen)	-	-0.006*** (0.001)	-0.006*** (0.002)	-0.008*** (0.002)
obs.		4,838	4,838	4,838	4,838
McFadden $\mathbf{R^2}$		-	0.0644	0.0204	0.0038
Predicted Probability		-	0.030 (0.002)	0.110 (0.004)	0.577 (0.007)

Notes: Rows 2-4 show the marginal effects of each attribute based on the logit model. The numbers in parentheses indicate standard errors. " *** " denotes statistical significance at the 1% level. Row 7 presents the average predicted probabilities based on the estimates from the logit model. Here, all instances of "AI" in the table refer to "pre-trained AI."

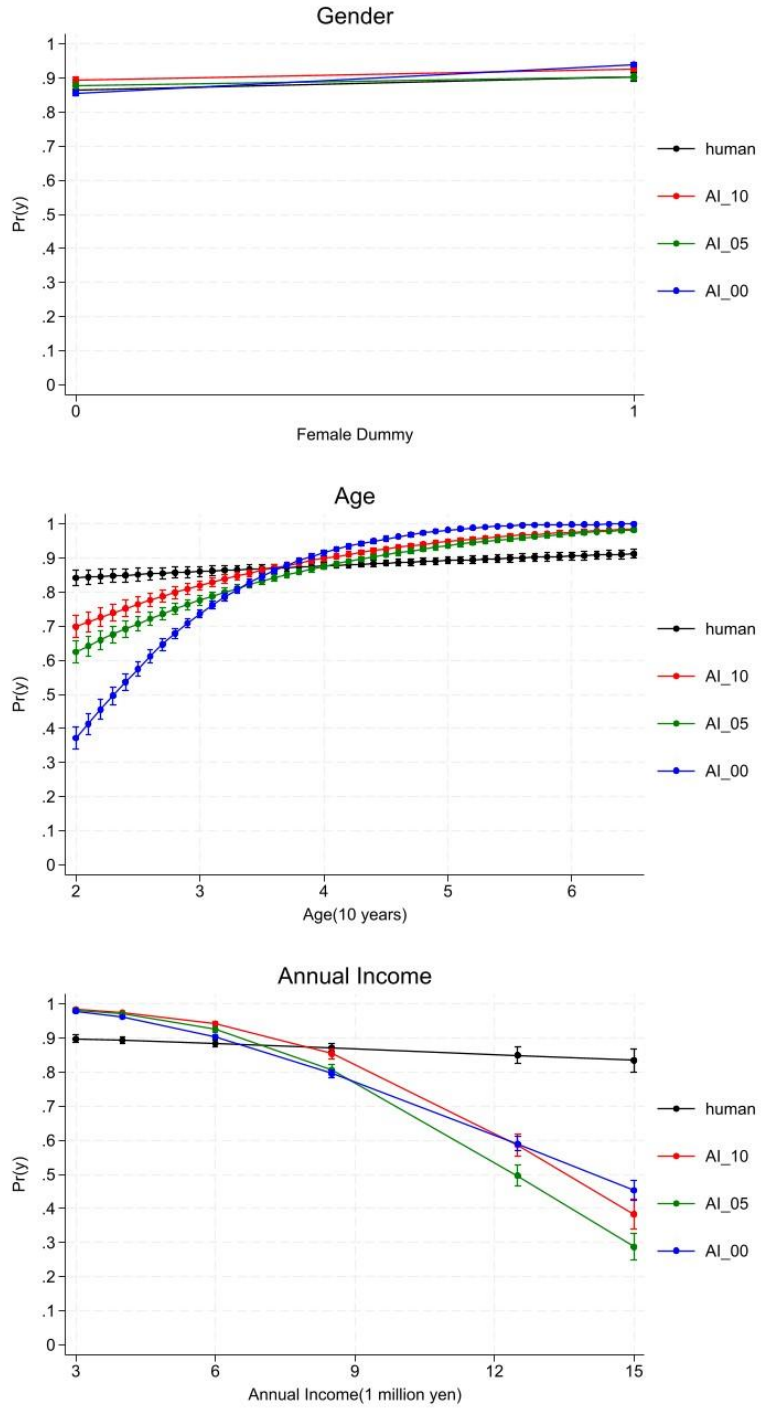

Figure A.3.1: Choice Probabilities: Humans vs. Pre-trained AI (Gain, 3 Attributes)

Notes: AI_10 refers to the responses from the pre-trained AI when queried with a temperature setting of 1.0. The same applies to AI_05 and AI_00. The vertical bars at each point in the figure represent 95% confidence intervals.

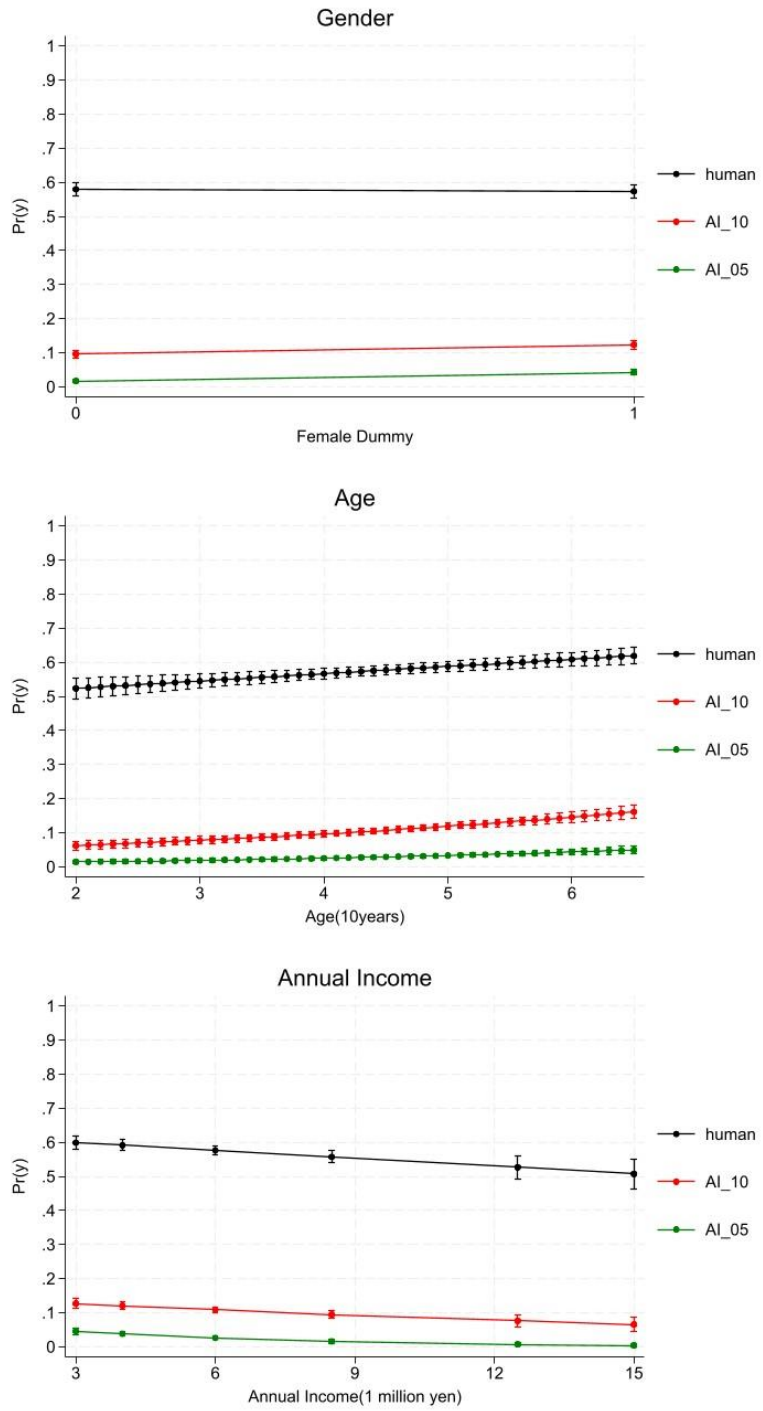

Figure A.3.2: Choice Probabilities: Humans vs. Pre-trained AI (Loss, 3 Attributes)

Notes: AI_10 refers to the responses from the pre-trained AI when queried with a temperature setting of 1.0. The same applies to AI_05 and AI_00. The vertical bars at each point in the figure represent 95% confidence intervals.

Table A.3.4: Logit Model Prediction Accuracy: Humans vs. Pre-trained AI

Question 1 Gain	AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	Human
Predicted value, Observed value				
(2, 2)	6.39% (n=309)	4.24% (n=205)	2.34% (n=113)	0% (n=0)
(2, 1)	1.53% (n=74)	2.00% (n=97)	1.34% (n=65)	0% (n=0)
(1, 2)	3.78% (n=183)	6.84% (n=331)	6.57% (n=318)	11.60% (n=561)
(1,1)	88.30% (n=4,272)	86.92% (n=4,205)	89.75% (n=4,342)	88.40% (n=4,277)
Prediction Accuracy	94.69%	91.15%	92.08%	88.40%

Question 2 Loss	AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	Human
Predicted value, Observed value				
(2, 2)	-	96.96% (n=4,691)	88.98% (n=4,305)	0.93% (n=45)
(2, 1)	-	3.04% (n=147)	11.02% (n=533)	1.03% (n=50)
(1, 2)	-	0% (n=0)	0% (n=0)	41.38% (n=2,002)
(1,1)	-	0% (n=0)	0% (n=0)	56.66% (n=2,741)
Prediction Accuracy	-	96.96%	88.98%	57.59%

Notes: All instances of "AI" in the table refer to the "pre-trained AI."

A.4 The Process of Fine-Tuning

We explain the process of fine-tuning. First, we describe the training data. Figure A.4.1 shows an excerpt from the JSONL format dataset used for fine-tuning. In fine-tuning with GPT, each training sample must consist of a sequence of 10 or more dialogues using three roles: "system", "user", and "assistant". The “system” defines the instructions the model should follow. The “user” provides the input or question to the model. The “assistant” provides the model’s expected response. In our dataset, we input virtual personal attributes (personas) under "system", questions shown in Section 3.2 under "user", and the choice value of respondents under "assistant", based on the classification of samples using the average values for each of the three attributes (gender, age, and income) (Table A.4). We create a total of 16 JSONL files, 8 for each of the two questions.

Figure A.4.2 shows the Python script used to train the GPT-4o-2024-08-06 model. Lines 1–3 of the script, “import ...”, read JSONL files and communicate with the OpenAI API. Line 9 specifies the path to the JSONL file containing analysis targets. Lines 12–18 read the JSONL training file and prepare it for uploading to the model. Lines 21–31 initiate the fine-tuning of GPT-4o-2024-08-06 with epoch=1 and LR=0.2. Lines 33–36 instruct the output of the progress status during fine-tuning[18]. Based on this script, AI can be fine-tuned using a method that applies the group K-fold cross-validation described in Chapter 4. The Python script used to present questions to fine-tuned AI and obtain responses is the same as the script in Figure A.2, but with the fine-tuned AI model name inserted on line 52.

[18] In fine-tuning, OpenAI system randomly shuffles the examples repeatedly, so the order of examples should not make a difference. Namely, each sample contributes independently to the learning.

{"messages": [{"role": "system", "content": "You are a 36-year-old man with an annual income of 3 million yen."}, {"role": "user", "content": " Imagine you receive an additional 50,000 yen on top of your current wealth, and are asked to choose between the following options. Which do you choose? Option 1: Lose 10,000 yen with certainty Option 2: A 50% chance of losing 20,000 yen, and a 50% chance of losing nothing. "}, {"role": "assistant", "content": "1"}]}
{"messages": [{"role": "system", "content": " You are a 57-year-old woman with an annual income of 3 million yen."}, {"role": "user", "content": " Imagine you receive an additional 50,000 yen on top of your current wealth, and are asked to choose between the following options. Which do you choose? Option 1: Lose 10,000 yen with certainty Option 2: A 50% chance of losing 20,000 yen, and a 50% chance of losing nothing."}, {"role": "assistant", "content": "2"}]}

Figure A.4.1: Dataset for Fine-Tuning

Notes: An excerpt from the JSONL file used for fine-tuning.

Table A.4: Classification Used to Create a Dataset for Fine-Tuning

Group	Gender	Age	Income	obs.(total: 4,838)
1	Female	Under average age (44.9 years)	Under average income (JPY 5.93 million)	657
2			Average income (JPY 5.93 million) or more	531
3		Average age (44.9 years old) or older	Under average income (JPY 5.93 million)	639
4			Average income (JPY 5.93 million) or more	626
5	Male	Under average age (44.9 years)	Under average income (JPY 5.93 million)	561
6			Average income (JPY 5.93 million) or more	540
7		Average age (44.9 years old) or older	Under average income (JPY 5.93 million)	565
8			Average income (JPY 5.93 million) or more	719

Notes: Based on the above classification, the 4,838 samples are divided into eight groups.

```python
import openai
import os
import json
# "***" denotes API key.
openai.api_key = "***"
# "***" denotes the file name of JSONL file uploaded.
file_path = "*****.jsonl"

# Uploading JSONL file to OpenAI
print("Uploading training file...")
response = openai.files.create(
    file=open(file_path, "rb"),
    purpose="fine-tune"
)
file_id = response.id
print(f"File uploaded successfully! File ID: {file_id}")

# Starting the fine-tuning process (using GPT-4o)
print("Starting fine-tuning process...")
fine_tune_job = openai.fine_tuning.jobs.create(
    training_file=file_id,
    model="gpt-4o-2024-08-06",
    hyperparameters={
        "n_epochs": 3,
        "learning_rate_multiplier": 0.02
    }
    )
fine_tune_id = fine_tune_job.id
print(f"Fine-tuning started! Job ID: {fine_tune_id}")
# Check the progress of fine-tuning
print("Checking fine-tuning status...")
status_response = openai.fine_tuning.jobs.retrieve(fine_tune_id)
print("Fine-tuning status:", status_response.status)
```

Figure A.4.2: Python Script for Fine-tuning

A.5 Summary of the Comparative Results with Other Models

Table A.5.1: Choice Results: Humans vs. Pre-trained AI (Model Comparison)

Question 1 Gain	GPT-4o	Gemini-2.5-flash	DeepSeek-v3	Human
Mean	0.911	0.798	0.806	0.884
Std. Dev.	0.285	0.402	0.396	0.320
Obs.	4,838	4,838	4,838	4,838

Question 2 Loss	GPT-4o	Gemini-2.5-flash	DeepSeek-v3	Human
Mean	0.110	0.337	0.543	0.571
Std. Dev.	0.313	0.473	0.498	0.495
Obs.	4,838	4,838	4,838	4,838

Notes: All instances of "AI" in the table refer to pre-trained AI. when queried with a temperature setting of 1.0.

Table A.5.2: Logit Model Estimation: Humans vs. Pre-trained AI (Gain, 3 Attributes; Model Comparison)

Question 1 Gain		GPT-4o	Gemini-2.5-flash	DeepSeek-v3	Human
Marginal Effects	Female dummy	0.033***	0.105***	0.081***	0.041***
		(0.007)	(0.008)	(0.009)	(0.009)
	Age (10 years)	0.057***	0.160***	0.136***	0.015***
		(0.003)	(0.003)	(0.004)	(0.003)
	Annual income (1 million yen)	-0.027***	-0.047***	-0.044***	-0.005***
		(0.001)	(0.001)	(0.001)	(0.001)
obs.		4,838	4,838	4,838	4,838
McFadden $\mathbf{R^2}$		0.3472	0.4652	0.3752	0.0140
Predicted Probability		0.911	0.798	0.806	0.884
		(0.004)	(0.004)	(0.004)	(0.005)

Notes: Rows 2-4 show the marginal effects of each attribute based on the logit model. The numbers in parentheses indicate standard errors. " *** " denotes statistical significance at the 1% level. Row 7 presents the average predicted probabilities based on the estimates from the logit model. Here, all instances of "AI" in the table refer to pre-trained AI. when queried with a temperature setting of 1.0.

Table A.5.3: Logit Model Estimation: Humans vs. Pre-trained AI (Loss, 3 Attributes; Model Comparison)

Question 2 Loss		GPT-4o	Gemini-2.5-flash	DeepSeek-v3	Human
Marginal Effects	Female dummy	0.028***	0.100***	0.018	-0.006
		(0.009)	(0.011)	(0.013)	(0.014)
	Age (10 years)	0.023***	0.122***	0.143***	0.021***
		(0.004)	(0.004)	(0.004)	(0.005)
	Annual income (1 million yen)	-0.006***	-0.052***	-0.011***	-0.008***
		(0.002)	(0.002)	(0.002)	(0.002)
obs.		4,838	4,838	4,838	4,838
McFadden $\mathbf{R^2}$		0.0204	0.2102	0.1315	0.0038
Predicted Probability		0.110	0.337	0.543	0.577
		(0.004)	(0.007)	(0.003)	(0.007)

Notes: Rows 2-4 show the marginal effects of each attribute based on the logit model. The numbers in parentheses indicate standard errors. " *** " denotes statistical significance at the 1% level. Row 7 presents the average predicted probabilities based on the estimates from the logit model. Here, all instances of "AI" in the table refer to pre-trained AI. when queried with a temperature setting of 1.0.

A.6 The Analysis of Risk Preference Questions in which the Monetary Stakes are Increased Tenfold

Question 1. Imagine you receive an additional JPY 300,000 on top of your current wealth and are asked to choose between the options below. Which option would you choose?

Option 1: Receive a guaranteed JPY 100,000

Option 2: A 50% chance of receiving 200,000 yen and a 50% chance of receiving nothing

Question 2. Imagine you receive an additional JPY 500,000 on top of your current wealth and are asked to choose between the options below. Which option would you choose?

Option 1: Lose 100,000 yen with certainty

Option 2: A 50% chance of losing JPY 200,000 and a 50% chance of losing nothing

Table A.6.1: Choice Results: Humans vs. Pre-trained AI (Monetary Stakes: Tenfold)

	Question 1 Gain		Question 2 Loss	
	Pre-trained AI temperature 1.0	Human	AI	Pre-trained AI temperature 1.0
Mean	0.883	0.888	0.157	0.619
Std. Dev.	0.321	0.315	0.364	0.486
Obs.	4,838	4,838	4,838	4,838

Table A.6.2: Logit Model Estimation: Humans vs. Pre-trained AI (Gain, 3 Attributes; Monetary Stakes: Tenfold)

Question 1 Gain		AI	Human
Marginal Effects	Female dummy	0.052***	0.038***
		(0.008)	(0.009)
	Age (10 years)	0.068***	0.013***
		(0.004)	(0.003)
	Annual income (1 million yen)	-0.033***	-0.004***
		(0.001)	(0.001)
obs.		4,838	4,838
McFadden $\mathbf{R^2}$		0.3370	0.0122
Predicted Probability		0.883	0.888
		(0.005)	(0.005)

Notes: Rows 2-4 show the marginal effects of each attribute based on the logit model. The numbers in parentheses indicate standard errors. " *** " denotes statistical significance at the 1% level. Row 7 presents the average predicted probabilities based on the estimates from the logit model. Here, all instances of "AI" in the table refer to pre-trained AI. when queried with a temperature setting of 1.0.

Table A.6.3: Logit Model Estimation: Humans vs. Pre-trained AI
(Loss, 3 Attributes; Monetary Stakes: Tenfold)

Question 2 Loss		AI	Human
Marginal Effects	Female dummy	0.064*** (0.010)	0.007 (0.014)
	Age (10 years)	0.034*** (0.004)	0.016*** (0.005)
	Annual income (1 million yen)	-0.007*** (0.002)	-0.006** (0.002)
obs.		4,838	4,838
McFadden $\mathbf{R^2}$		0.0301	0.0023
Predicted Probability		0.157 (0.005)	0.619 (0.007)

Notes: Rows 2-4 show the marginal effects of each attribute based on the logit model. The numbers in parentheses indicate standard errors. " *** ", and " ** " denotes statistical significance at the 1%, and 5% level, respectively. Row 7 presents the average predicted probabilities based on the estimates from the logit model. Here, all instances of "AI" in the table refer to pre-trained AI. when queried with a temperature setting of 1.0.