NBER WORKING PAPER SERIES

BEHAVIORAL ECONOMICS OF AI:
LLM BIASES AND CORRECTIONS

Pietro Bini
Lin William Cong
Xing Huang
Lawrence J. Jin

Working Paper 34745
http://www.nber.org/papers/w34745

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2026

We are grateful to Nicholas Barberis, James Choi, Clifton Green, William Goetzmann, David Hirshleifer, Gerard Hoberg, Camelia Kuhnen, Devin Shanthikumar, Kelly Shue, Siew Hong Teoh, Luyao Zhang, and seminar participants at the Federal Reserve Board, the University of California Los Angeles, the 2024 Conference for Financial Economics and Accounting, the 2025 Conference on Emerging Technologies in Accounting and Financial Economics, the 2025 Chicago Booth Conference on Behavioral Approaches to Financial Decision Making, and the 2026 AFA Annual Meeting for helpful discussions and comments. Jordan Velte and Shuhuai Zhang provided excellent research assistance. Bini and Cong acknowledge financial support from Ripple's University Blockchain Research Initiative. Please send correspondence to Jin at lawrence.jin@cornell.edu or Cong at will.cong@cornell.edu. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Pietro Bini, Lin William Cong, Xing Huang, and Lawrence J. Jin. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Behavioral Economics of AI: LLM Biases and Corrections
Pietro Bini, Lin William Cong, Xing Huang, and Lawrence J. Jin
NBER Working Paper No. 34745
January 2026
JEL No. D03, G02, G11, G4, G40, G41

ABSTRACT

Do generative AI models, particularly large language models (LLMs), exhibit systematic behavioral biases in economic and financial decisions? If so, how can these biases be mitigated? Drawing on the cognitive psychology and experimental economics literatures, we conduct the most comprehensive set of experiments to date—originally designed to document human biases—on prominent LLM families across model versions and scales. We document systematic patterns in LLM behavior. In preference-based tasks, responses become more human-like as models become more advanced or larger, while in belief-based tasks, advanced large-scale models frequently generate rational responses. Prompting LLMs to make rational decisions reduces biases.

Pietro Bini
Boston University
Questrom School of Business
Finance
pbini@bu.edu

Lin William Cong
Cornell University
and NBER
will.cong@cornell.edu

Xing Huang
Cornell University
Department of Finance
and Washington University in St Louis
xing.huang@cornell.edu

Lawrence J. Jin
Cornell University
and NBER
lawrence.jin@cornell.edu

A internet appendix is available at http://www.nber.org/data-appendix/w34745

1. Introduction

Artificial intelligence (AI), especially generative large language models (LLMs), is becoming increasingly essential in daily work and general economic activities. For example, banks and FinTech firms are integrating generative AI (GenAI) technologies into operations management, customer service, financial advice, and risk assessment (Vidal, 2023; Tomlinson, Laughridge, and Dockar, 2024). Researchers are exploring the potential for LLMs to enhance experimentation that studies human behavior (Charness, Jabarian, and List, 2023; Korinek, 2023; Bail, 2024). However, much remains unknown about how AI algorithms and agents behave systematically, particularly in economic and financial decisions, or whether their behavior resembles human behavior. Understanding the "behavioral economics" of AI—potentially a novel class of agents (Tegmark, 2017)—starting with LLMs is urgent for assessing and improving the technology's utility, safety, and appropriateness.

Recent studies have begun to examine the reliability of LLMs in expectation formation and decision making, with a focus on the behavior of ChatGPT.[1] Our paper not only adds to this work but also conceptually introduces a new field—the behavioral economics of AI—by establishing benchmark results: we conduct the most comprehensive set of experiments to date, originally designed to document human biases but now applied to investigate the biases of multiple prominent LLM families; we systematically compare LLM responses with both rational and human responses; and we explore methods for correcting these biases. An important goal of the paper is to develop a public database of experimental questions for ongoing evaluations of behavioral biases across various LLMs.

We start by exploring two broad approaches for conducting experiments that allow us to document the behavioral biases of LLMs. The first approach draws on the cognitive psychology literature, originating with Ellsberg (1961) and Kahneman and Tversky (1973, 1979), which uses carefully designed experimental questions to assess psychological biases in humans. From this literature, we select a comprehensive set of experiments covering both questions that study the psychology of preferences and questions that study the psychology of beliefs. Our selection ensures

---

[1] For example, ChatGPT's behavior has been examined in both individual settings (Chen et al., 2023; Ma, Zhang, and Saunders, 2023; Chen et al., 2024; Bybee, 2025; Hansen et al., 2025) and game-theoretic settings (Bauer et al., 2023; Mei et al., 2024; Fan et al., 2024; Brookins and DeBacker, 2024; Manning and Horton, 2025).

the inclusion of experiments that document the psychological biases that are first-order important in financial markets.[2] For each question, we design a prompt suitable for LLMs, hence allowing us to elicit responses from these models and analyze their behavior. The second approach draws on recent experimental economics studies, which, compared to the cognitive psychology literature, feature experimental tasks that are more closely tied to economic and financial settings. We adapt these tasks for LLMs to investigate the behavioral biases they exhibit in financial decision making.

With the experimental questions in hand, we collect responses through an application programming interface (API) from four prominent families of LLMs: OpenAI's ChatGPT, Anthropic Claude, Google Gemini, and Meta Llama.[3] For each family, we consider two variations. First, we examine an advanced version of the model alongside an older version; this allows us to study time-series variation in the model's degree of behavioral biases. Second, for the advanced model, we compare a version with a large parameter scale to one with a smaller scale; this allows us to study cross-sectional variation in the model's degree of behavioral biases.

Analyzing the responses reveals five observations. First, for questions from the cognitive psychology literature that document biases in preferences, the LLMs' answers exhibit a clear pattern: as models become more advanced or larger, responses become increasingly human-like and they are irrational according to the Expected Utility framework. For example, Claude 3 Opus, an advanced large-scale LLM, answers four out of six preference-based questions in a way that is consistent with human responses. In comparison, Claude 3 Haiku, another advanced model with a smaller scale, gives human-like answers to three out of the six questions, while Claude 2, an older version, gives human-like answers only to one out of the six questions.

Second, for questions from the cognitive psychology literature that document biases in beliefs, the LLMs' answers exhibit the *opposite* pattern: more advanced or larger models produce increasingly rational responses. For example, Gemini 1.5 Pro, a highly advanced large-scale LLM, answers all ten belief-based questions correctly. By contrast, Gemini 1.5 Flash, another advanced but smaller model, answers five questions correctly, while Gemini 1.0 Pro, an older version, answers only two questions correctly. Overall, three of the four advanced large-scale LLMs we examine—GPT-4, Claude 3 Opus, and Gemini 1.5 Pro—produce predominantly rational answers to belief-based

---

[2] Barberis (2018) argues that prospect theory preferences, overextrapolation, and overconfidence are the three main psychological biases that drive investor behavior, firm behavior, and asset prices in financial markets.

[3] These were the extant LLMs when we begun our study in 2023.

questions.

Third, substantial heterogeneity emerges when we compare responses across LLM families. For preference-based questions, Gemini's responses are less rational and more human-like compared to those from ChatGPT, while the responses from Claude or Llama are generally similar to those from ChatGPT. For belief-based questions, Meta Llama's responses are less rational and more human-like compared to those from GPT, while the responses from Anthropic Claude or Google Gemini are similar to those from GPT.

Fourth, we examine LLMs' responses to questions from two recent experimental economics studies. Afrouzi et al. (2023) ask human participants to observe a sequence of past realizations of a random variable and then forecast its future values; the random variable follows an autoregressive process. We elicit LLMs' responses in this setting and find that advanced small-scale LLMs—GPT-4o, Claude 3 Haiku, and Gemini 1.5 Flash—produce forecasts that are human-like and irrational: they perceive an autoregressive process that is more persistent than the true process. By contrast, their larger-scale counterparts generate more rational forecasts, with perceived persistence similar to the true persistence.

Bose et al. (2022) present human participants with stock price trajectories and ask them how much to invest in each stock. Replicating this setting for LLMs, we find that large-scale models—GPT-4, Claude 3 Opus, and Gemini 1.5 Pro—make investment decisions that are more human-like than their smaller-scale counterparts: investment depends more strongly on the visual salience of a stock's price trajectory, a preference-based variable identified by Bose et al. (2022) as driving human investment behavior. Taken together, these results suggest that the patterns documented with cognitive psychology questions also hold in experimental economics studies: for preference-based questions, larger models produce increasingly human-like responses, while for belief-based questions, larger models produce increasingly rational responses.[4]

Finally, we explore methods for correcting the observed behavioral biases. Among the methods we test, one seems effective while the others are not. The effective method involves a brief role-priming instruction that asks an LLM to think of itself as a rational investor who makes decisions using the Expected Utility framework; this instruction is provided before the LLM answers any

---

[4] For the Afrouzi et al. (2023) and Bose et al. (2022) tasks, LLMs must process graphical inputs. Six of the twelve LLMs we examine do not support graphical inputs, so the tasks are not run on these models. See Section 2.3 for a detailed discussion.

question. We find that, for both preference-based and belief-based questions, this role-priming instruction makes LLM responses more rational and less human-like, although the magnitude of improvement is economically modest. We also find that role priming affects LLM responses through the confidence levels LLMs assign to their choices as well as their self-reported reasoning type, where type A corresponds to intuitive thinking and type B corresponds to analytical thinking and calculations. The other methods, which combine the role-priming instruction with additional genuinely useful information, are found to be ineffective in reducing biases. Overall, these results indicate that debiasing LLMs remains a challenging task.

The five observations stated above are descriptive, yet informative. Although a full analysis of the mechanisms driving these results is beyond the scope of the paper, we put forth two conjectures. First, why do more advanced or larger-scale models become more human-like when responding to preference-based questions? We conjecture this is partly because advanced, large-scale LLMs increasingly rely on Reinforcement Learning from Human Feedback (RLHF), a training process that aligns the underlying model with human preferences as reflected in human feedback (Stiennon et al., 2020). Second, why do more advanced or larger-scale models become more rational when responding to belief-based questions? We conjecture this is partly because the larger training data and greater computational capacity of advanced, large-scale LLMs enable them to better identify statistical ground truths based on which they respond to belief-based questions. Investigating these conjectures could provide guidance for the design of future LLMs.

Literature. Over the past five decades, the cognitive psychology literature (Ellsberg, 1961; Kahneman and Tversky, 1973, 1979; Tversky and Kahneman, 1981; Rapoport and Budescu, 1992, 1997; Frederick, Loewenstein, and O'Donoghue, 2002; Barberis and Thaler, 2003) and the experimental economics literature (Lian, Ma, and Wang, 2018; Bose et al., 2022; Afrouzi et al., 2023) have systematically documented behavioral biases exhibited by *human* participants. A related strand of research focuses on developing methods to mitigate these human biases (Choi et al., 2004; Thaler and Sunstein, 2008; DellaVigna and Linos, 2022). In this paper, we extend these research fields by studying the behavioral economics of AI—doing so is important for addressing two fundamental questions.

The first question concerns the use of LLMs as research tools for studying human behavior. As

GenAI advances, LLMs are increasingly used for social science research: recent studies highlight LLMs' potential to enhance research design, experimentation, data analysis, and agent-based modeling of complex activities (Charness, Jabarian, and List, 2023; Korinek, 2023; Bail, 2024). These studies generally treat LLMs as neutral, unbiased tools. We challenge this assumption by systematically investigating LLM behavior using insights from cognitive psychology and experimental economics. We document LLM biases and their heterogeneity across preference-based and belief-based questions, model versions, and scales. Understanding these biases is critical for evaluating the reliability of LLM-based experiments and simulations.

The second question concerns the behavior of AI agents in tasks traditionally performed by humans. These agents are increasingly deployed in economic and financial settings, yet the reliability of their performance remains unclear. Recent work shows mixed patterns: Chen et al. (2023) find that GPT-3.5 Turbo exhibits higher economic rationality and lower choice heterogeneity than humans in multiple domains of individual decision making; Mei et al. (2024) show that GPT-4 exhibits human-like traits in games; Chen et al. (2025) document biased beliefs in LLM forecasts of stock returns that are commonly observed among humans; Bowen et al. (2025) show that LLMs exhibit strong racial biases in mortgage underwriting, which can be mitigated with prompts that require unbiased decisions; Cook et al. (2025) find that, in allocation games, LLM preferences exhibit inequality aversion and are malleable under interventions; and Ouyang, Yun, and Zheng (2024) study how LLM risk preferences in financial settings can be aligned with human ethical standards. Together, these studies suggest that LLM behavior sometimes mirrors human behavior but is sensitive to prompt framing, training data, and model architecture.

Compared with the studies mentioned above and the broader literature on LLM performance and algorithmic biases, our work is among the first to advocate a systematic exploration of the behavioral economics of AI, treating GenAI agents as a novel class of economic agents. We advance the literature in several ways. First, we analyze multiple prominent LLM families and examine both cross-sectional and time-series variations within each family.[5] Second, we document biases across both preference-based and belief-based questions, drawing on the cognitive psychology and experimental economics literatures, and covering behavioral biases that are first-order important

---

[5] Our work is contemporaneous with Chen et al. (2023), Mei et al. (2024), Ouyang et al. (2024), Chen et al. (2025), and Bowen et al. (2025). Nonetheless, a longer data sample is needed to study time-series variation in LLM responses.

in financial markets.[6] Third, we investigate methods to mitigate these biases, comparing different approaches and propose new ones. Overall, our work lays the foundation for systematically documenting LLM biases and exploring debiasing methods, contributing to the emerging literature on LLM evaluations.[7]

The rest of the paper proceeds as follows. Section 2 discusses the experimental design. Section 3 presents our results on LLM responses to preference-based and belief-based questions. Section 4 explores methods for correcting the observed behavioral biases of LLMs, and Section 5 concludes.

2. Experimental Design

This section describes the experimental design. We first discuss the selection of questions that study either the psychology of preferences or the psychology of beliefs. We then describe the selection of LLMs. Finally, we outline the design of API prompts that allow us to systematically elicit LLM responses to the experimental questions.

2.1. Selection of Experimental Questions

Traditional theories in economics and finance posit that economic agents make rational decisions. Here, rationality has two components. The first is rational preferences, namely that agents make decisions according to the Expected Utility framework proposed by Von Neumann and Morgenstern (1944). The second is rational beliefs, namely that agents incorporate new information into their beliefs according to Bayes' law.

While traditional theories serve as a rational benchmark for economic studies, decades of research from cognitive psychology cast doubt on such theories. Specifically, through carefully designed experimental questions, the psychology literature has documented *actual* behaviors of human participants that systematically deviate from rational decision making. To illustrate, consider the following question posed to human participants by Kahneman and Tversky (1979):

---

[6] Our approach is consistent with Binz and Schulz (2023) and Shiffrin and Mitchell (2023): treating an LLM as a participant in a psychology experiment and studying its responses can help understand its mechanisms of reasoning and decision making.

[7] Recent work by Vafa, Rambachan, and Mullainathan (2024) finds that many LLMs, in particular the highly capable models such as GPT-4, perform poorly on tasks that humans expect them to perform well; this discrepancy highlights the necessity of systematic LLM evaluations. See Chang et al. (2024) for a review of LLM evaluations across multiple domains.

> "In addition to whatever you own, you have been given 1,000. You are now asked to choose between A: (1,000, .50), and B: (500)."

Here, (1,000, .50) means winning $1,000 with 50% probability and winning zero with 50% probability, while (500) means winning $500 with certainty. For this question, the majority of participants would choose option B. The same set of participants are then asked a separate question:

> "In addition to whatever you own, you have been given 2,000. You are now asked to choose between C: (–1,000, .50), and D: (–500)."

Here, (–1,000, .50) means losing $1,000 with 50% probability and losing zero with 50% probability, while (–500) means losing $500 with certainty. For this question, the majority of participants would choose option C.

It is easy to verify that, in terms of monetary payoffs, option A from the first question is equivalent to option C from the second question, and option B from the first question is equivalent to option D from the second question. As a result, the same participant choosing option B from the first question and option C from the second question is a clear violation of the Expected Utility framework.

Through experimental questions such as the one described above, cognitive psychologists have carefully examined human psychology of preferences—including both risk and time preferences—and human psychology of beliefs, and they have documented a comprehensive set of behavioral biases. In this paper, we ask LLMs to answer the same experimental questions and collect their responses through the prompt design described in Section 2.3; in other words, we replace human participants by LLMs. This approach allows us to systematically document the behavioral biases of LLMs and compare their behavior with human behavior. Table 1 provides a summary of all the cognitive psychology questions studied in this paper.

[Place Table 1 about here]

Two observations are worth noting. First, for each question in Table 1, an LLM response can be classified into one of three categories: a rational response that is derived from rational preferences and rational beliefs; a human-like (irrational) response that corresponds to the response from the majority of human participants; or a non-human-like response, which is neither rational nor human-like. Second, Table 1 includes experimental questions that are designed to document prospect theory preferences (Questions 1 to 3), overextrapolation (Questions 7 to 10), and overconfidence (Questions 15 and 16). These three psychological biases, according to Barberis (2018), are the "big three" biases that are of first-order importance for understanding investor behavior, firm behavior, and asset prices in financial markets.

Compared to the cognitive psychology literature, the experimental economics literature examines human behavior by designing and conducting experimental tasks that are more closely tied to real-world economic and financial settings. To broaden the scope of our analysis, we also collect LLM responses to a set of tasks from this literature, focusing on two recent works: those of Afrouzi et al. (2023), which study investor beliefs, and those of Bose et al. (2022), which study investor preferences.

We begin with the experiments of Afrouzi et al. (2023), in which human participants first observe a sequence of past realizations of a random variable x_t and then forecast its future values; the time-series evolution of this random variable is governed by the following autoregressive process:

$$x_t = \mu + \rho x_{t-1} + \epsilon_t, \tag{1}$$

where ρ measures the persistence of the process and ϵ_t is an i.i.d. Gaussian shock.

We elicit LLM responses in this setting. As in Afrouzi et al. (2023), we consider three experiments. In the baseline experiment, an LLM is endowed with the knowledge that the evolution of x_t follows a "stable random process." The LLM first observes 40 past realizations of x_t, ranging from x_1 to x_{40}, and is then asked, at time 40, to forecast the next two outcomes, x_{41} and x_{42}. Subsequently, it observes the realization of x_{41} and is asked, at time 41, to forecast x_{42} and x_{43}. This procedure continues until the LLM observes 44 past realizations of x_t and is asked, at time 44, to forecast x_{45} and x_{46}.

The second and third experiments are variants of the baseline. In the second experiment, at each time t, the LLM is asked to forecast x_{t+1} and x_{t+5}; for example, at time 40, it observes x_1 to x_{40} and then forecasts x_{41} and x_{45}. In the third experiment, the LLM is endowed with more detailed knowledge that the evolution of x_t follows "a fixed and stationary AR(1) process: $x_t = \mu + \rho x_{t-1} + \epsilon_t$, with a given μ, a given ρ in the range [0,1], and an ϵ_t that is an i.i.d. random shock."

For each experiment and a wide range of ρ values, we compare the true ρ with $\hat{\rho}$, the "perceived" autoregressive coefficient implied by LLM forecasts. This comparison allows us to document biases in LLM beliefs through experiments that mimic real-world forecasting tasks.

We next turn to the experiment of Bose et al. (2022), in which human participants first observe a stock price trajectory and then makes an investment decision: assuming an endowment of 1,000 monetary units, they indicate how much to invest in the stock over a 12-month period. We replicate this setup with LLMs by eliciting their responses in the same setting. We then assess whether LLM responses are human-like by examining whether the same preference-based factors that drive investment decisions of human participants also explain the LLMs' decisions. This analysis allows us to understand LLM preferences through experiments that mimic real-world investment tasks.

2.2. Selection of LLMs

We select twelve LLMs from four prominent families of Generative Pre-trained Transformers (GPT): ChatGPT, Anthropic Claude, Google Gemini, and Meta Llama. For each family, we select three models: a benchmark model defined as the most recent and best-performing one available at the time of writing, its smaller-scale version, and its predecessor. For ChatGPT, we use GPT-4 as the benchmark, GPT-4o as the smaller-scale version, and GPT-3.5 Turbo as the predecessor. For Anthropic Claude, we use Claude 3 Opus as the benchmark, Claude 3 Haiku as the smaller-scale version, and Claude 2 as the predecessor. For Google Gemini, we use Gemini 1.5 Pro as the benchmark, Gemini 1.5 Flash as the smaller-scale version, and Gemini 1.0 Pro as the predecessor. Finally, for Meta Llama, we use Llama 3 70B as the benchmark, Llama 3 8B as the smaller-scale version, and Llama 2 70B as the predecessor. Table 2 summarizes all the LLMs we examine.

[Place Table 2 about here]

These twelve models differ both across and within families, particularly along three dimensions: the size of the training data, the design of the model architecture, and the reinforcement learning algorithm. In terms of training data, newer models are generally trained on larger datasets. For example, Meta Llama reports that the training dataset of Llama 3 consists of over 15 trillion tokens, while Llama 2 consists of 1.8 trillion tokens only.[8]

In terms of model architecture, we note three differences across models. First, model specifics such as the context window—the maximum number of words that a model can take as input—and the number of parameters vary widely from one model to another. For example, among older models, Claude 2 has an estimate of 200 billion parameters and a context window of approximately 100,000 tokens, whereas Llama 2 has 70 billion parameters and a context window of approximately 4,000 tokens. Second, within each family, model architectures have evolved significantly across the two generations that we consider. In particular, for ChatGPT, Anthropic Claude, and Google Gemini, the most significant evolution is the transition from a single-transformer architecture to a multi-transformer mixture-of-experts architecture.[9] Third, within each family and each generation, model architecture may differ between the benchmark model and its smaller-scale version. The smaller versions are often obtained by applying compression techniques to the benchmark model; for example, Gemini 1.5 Flash is a distilled version of Gemini 1.5 Pro.[10,11]

In terms of reinforcement learning algorithms, each model relies on a different implementation of the Reinforcement Learning from Human Feedback (RLHF) algorithm to align outputs with human preferences. For example, Anthropic Claude combines RLHF with Constitutional AI, a method designed to align model behavior with human principles of helpfulness, harmlessness, and honesty (Bai et al., 2022).

2.3. Prompt Design

We collect LLM responses to each experimental question through an application programming interface (API). The API takes as input a "prompt," which is a text file submitted to an LLM

---

[8] For more details about the characteristics of Meta Llama 3, see: https://ai.meta.com/blog/meta-llama-3/. For the other three LLM families, the exact training data are not publicly disclosed; nonetheless, newer models in general tend to be trained on more data. Brown et al. (2020) report that OpenAI used around 500 billion tokens to train GPT-3.5. The exact number of tokens used to train GPT-4 is unknown, although unofficial sources suggest around 13 trillion tokens; see https://semianalysis.com/2023/07/10/gpt-4-architecture-infrastructure/.

[9] Mixture-of-experts architectures use a "router," or gating network, to activate specific experts for each input token (Shazeer et al., 2017). The sparsity that arises from activating only a fraction of parameters for each input enables better scaling. For example, GPT-3.5 Turbo uses a single-expert architecture with 175 billion parameters, while unofficial sources suggest that GPT-4 uses a mixture-of-experts architecture that consists of multiple transformers with approximately 110 billion parameters each for a total of over 1 trillion parameters.

[10] Common compression techniques include quantization, which reduces parameter precision, and pruning, which removes less important connections from the neural network.

[11] In contrast, Llama 3 8B and Llama 3 70B share similar architectures and training data but differ in scale (8 billion versus 70 billion parameters).

to elicit a response. Below, we describe the prompt design that allows for elicitation of desired responses from LLMs.

A proper prompt needs to satisfy two requirements. First, it should instruct the LLM to provide standardized responses suitable for subsequent analysis. Second, it should phrase questions in a manner that is comparable to the original experimental questions used to study human behavior. Given these requirements, Fig. 1 provides an example: the prompt we use to elicit LLM responses to a question that Kahneman and Tversky (1979) designed to document diminishing sensitivity as a key element of prospect theory.

[Place Fig. 1 about here]

Fig. 1 shows that the prompt is structured in three parts; this structure applies to all experimental questions listed in Table 1. The first part contains general instructions that ask the LLM to consider a specific set of experimental scenarios; in Fig. 1, this part begins with “Instructions” and ends with “completely separate from the other.” two lines below. The second part contains a code block that instructs the LLM to format its responses in a standardized JSON format; in Fig. 1, this part begins with “The output should be” and ends with “} ‘‘‘”.[12] The third part is the main element of the prompt. It contains the precise experimental questions originally designed by psychologists to study human behavior; in Fig. 1, this part begins with “Scenario A:” and ends with “calculations).” from Scenario B. At the end of each scenario, additional instructions are given to the LLM, to ensure that it provides the specific set of responses we elicit.

Two observations are worth noting. First, under our prompt design, an LLM response typically contains, for each experimental scenario, four components: choice, confidence, explanation, and reasoning. Here, “choice” refers to the explicit decision made by the LLM—for example, whether it accepts or turns down a risky gamble.[13] “Confidence” refers to the confidence level the LLM assigns to its choice, measured on a score between 0 and 1. “Explanation” refers to a brief explanation that the LLM provides to justify its choice. And “reasoning” requires the LLM to select between

---

[12] This part of the prompt requires formatting LLM responses as a snippet that contains a JSON object within a code block. JSON is a widely used format that stores data as key-value pairs. Encapsulating the JSON object within a code block helps ensure that LLM responses adhere to the pre-specified format.

[13] Instead of eliciting a “choice” among multiple options, Question 10 (regarding “base rate neglect”) and Question 12 (regarding “gambler’s fallacy”) ask for an estimate of a probability, while Question 14 (regarding “anchoring”) asks for an estimate of a percentage number.

two reasoning types: type “A,” which corresponds to reasoning that is based more on intuitive thinking, and type “B,” which corresponds to reasoning that is based more on analytical thinking and calculations.

Second, many of the experimental questions we examine document behavioral biases by eliciting responses from the *same* participant across different scenarios. For example, as discussed in Section 2.1, Kahneman and Tversky (1979) document diminishing sensitivity by having the same human participant answer two different questions—one that frames lottery payoffs as gains and another that frames them as losses. Such experimental questions require a within-subject design, which allows us to treat an LLM as a participant and elicits its responses across different scenarios. To implement this design, we combine multiple questions into a single API call; we treat each API call as an individual participant; and we include in the prompt an instruction that asks the LLM to “treat each scenario as completely separate from the other.”[14] The Internet Appendix provides the prompt designs for all sixteen questions listed in Table 1.

The above discussion is concerned with the prompt design that implements experimental questions from the cognitive psychology literature. We conclude this section by making three observations about two separate prompts, each corresponding to one of the experiments: the Afrouzi et al. (2023) experiments and the Bose et al. (2022) experiment described in Section 2.1. First, these experiments require not only textual inputs but also *graphical* inputs: participants are presented with textual instructions and figures that plot either past realizations of a random variable or stock price trajectories. To satisfy this requirement, an LLM must support graphical inputs; this leads to the exclusion of six LLM platforms.[15] For the remaining LLMs, we follow platform-specific guidelines when uploading figures.[16] Second, LLMs do not always provide precise forecasts or investment decisions when presented with figures; sometimes, they refuse to respond. To address this issue, for both the Afrouzi et al. (2023) and Bose et al. (2022) experiments, we include the following sentences in the instruction: “For the following question, please provide an estimate to the best of your knowledge. Please ensure that you always provide a concrete numerical answer when prompted to

---

[14] LLM responses can be stochastic—identical questions posed multiple times to the same model can yield different responses. We therefore view each API call as an individual participant.

[15] All three Meta Llama models—Llama 3 70B, Llama 3 8B, and Llama 2 70B—as well as GPT-3.5 Turbo, Claude 2, and Gemini 1.0 Pro do not support graphical inputs.

[16] Google Gemini directly processes figures uploaded as .jpg files. ChatGPT and Anthropic Claude, however, require encoding a binary image into bytes, which are then converted into a regular UTF-8 string format.

do so.” Third, the Afrouzi et al. (2023) experiments require that the same individual makes multiple rounds of forecasts, with each round depending on the textual and graphical inputs presented up to that point in time. To enforce this sequential dependence, we implement a sequence of API calls. In particular, for each call, we feed the entire conversation history—including all previous prompts and responses—into the LLM, thereby preserving the structure of the original experiments.

3. Behavioral Biases of LLMs

We now document patterns in LLM responses to the experimental questions drawn from the cognitive psychology literature and the experimental economics studies. We begin with a baseline analysis of the four highly advanced large-scale LLMs, which we treat as our benchmark models. We analyze how they respond to the cognitive psychology questions, with a focus on whether these models are more likely to produce rational or human-like responses. A central feature of this analysis is to draw distinction between LLM responses to preference-based questions and their responses to belief-based questions. We then explore heterogeneity in LLM responses across LLM families, model generations, and model scales. Finally, we examine LLM responses to questions from the experimental economics tasks, which are more closely tied to real-world economic and financial decision making.

3.1. Baseline Results

This section presents our baseline results. We first describe the procedure for data collection. We then analyze the responses of the four benchmark models—GPT-4, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3 70B—to the sixteen experimental questions drawn from the psychology literature, as listed in Table 1.

For each question and each model, we collect 100 responses; that is, for each LLM, we iterate over each question 100 times. Each iteration consists of an API call submitted to the model, with the prompt for the specific question provided as input along with a key temperature parameter. This parameter controls the randomness of the model. For our baseline analysis, we set the temperature parameter to 0.5, the recommended value for most LLM families.[17] Note that setting the

---

[17] The range of the temperature parameter varies across platforms: for ChatGPT and Anthropic Claude, it is [0, 1]; for Meta Llama, it is [0, 5]; for Gemini 1.0 Pro, it is [0, 1]; and for Gemini 1.5 Pro and Gemini 1.5 Flash, it is [0,

temperature parameter to zero results in deterministic outputs, while higher values increase the randomness of LLM responses.[18]

We collect and analyze each LLM response, categorizing it into one of three groups: rational, human-like, or other. A response is categorized as rational if the LLM's choice or estimate aligns with that of an agent who has rational preferences and rational beliefs; it is categorized as human-like if it is irrational but aligns with the most common behavior observed in human participants from prior psychology research; and it falls into the "other" category if it is neither rational nor human-like. Take the diminishing sensitivity question from Fig. 1 as an example. A rational response, according to the Expected Utility framework, is to choose option B, the option that indicates risk aversion, in both Scenario A and Scenario B. Kahneman and Tversky (1979) show that the majority of human participants choose option B in Scenario A and option A in Scenario B; if an LLM makes the same choices, the response is categorized as human-like. If, however, the LLM selects option A in Scenario A and option B in Scenario B, or selects option A in both scenarios, the response is categorized as "other."

[Place Fig. 2 and Table 3 about here]

Fig. 2 summarizes responses from the four benchmark models of GPT-4, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3 70B. For each model, the sixteen experimental questions from cognitive psychology are divided into two groups: preference-based questions (left panel) and belief-based questions (right panel). The results are presented using bar charts that show the proportion of responses categorized as rational (blue), human-like (red), or other (gray).[19] Table 3 presents the same results in tabular form and includes a binomial test for each question, where the null hypothesis states that the proportion of rational (or human-like) responses is less than or equal to 50%.

2].

[18] Specifically, during the iterative process of generating each word (token) in a response, the LLM first forms a probability distribution over all possible tokens in its dictionary and then draws the next token from this distribution. The temperature parameter reshapes the distribution: higher values make it more uniform, therefore increasing the randomness of the output token. Two other parameters, k and p, also affect the selection of the output token: top-k sampling restricts the selection to the top-k most probable tokens, while top-p sampling retains a subset of the top-k tokens whose cumulative probability, when normalized by the total probability of the top-k tokens, exceeds the threshold p. In our analysis, we set k to its default value of 50 and p to its default value of 0.9.

[19] Fig. IA.1 and Fig. IA.2 in the Internet Appendix present the proportion of rational, human-like, or other responses for advanced small-scale models and older models.

Two observations are worth emphasizing. First, the majority of LLM responses fall into either the rational category or the human-like category, with the responses classified as “other” in just a few cases. For GPT-4, such “other” responses arise only for Question 3, which pertains to the probability weighting element of prospect theory. For Claude 3 Opus, “other” responses appear in two preference-based questions—Question 3 on probability weighting and Question 4 on narrow framing—and two belief-based questions—Question 10 on base rate neglect and Question 15 on overprecision. For Gemini 1.5 Pro, “other” responses occur only in Question 3. Finally, for Llama 3 70B, “other” responses are observed in one preference-based question—Question 3 on probability weighting—and one belief-based question—Question 7 on sample size neglect.

Second, a comparison between the left and right panels of Fig. 2 reveals a clear pattern: LLM responses to preference-based questions tend to be more human-like, whereas their responses to belief-based questions tend to be more rational. Table 3 confirms this result. For a large fraction of preference-based questions, a binomial test confirms, with a confidence level greater than 99%, that LLMs produce human-like responses more than 50% of the time. Specifically, Gemini 1.5 Pro has the majority of its responses categorized as human-like in five out of six questions; Claude 3 Opus has the majority of its responses categorized as human-like in four out of six questions; and GPT-4 and Llama 3 70B have the majority of their responses categorized as human-like in three out of six questions. By contrast, for most belief-based questions, LLMs produce rational responses more than 50% of the time. Specifically, Gemini 1.5 Pro has the majority of its responses categorized as rational in all ten questions; both GPT-4 and Claude 3 Opus have the majority of their responses categorized as rational in eight out of ten questions; and Llama 3 70B has the majority of its responses categorized as rational in five out of ten questions.

3.2. Heterogeneity in LLM Responses

While Section 3.1 documents systematic patterns in LLM responses for the four benchmark models, we now broaden the analysis to include all twelve models. Within each LLM family, we consider three types of models: a highly advanced, large-scale benchmark model; a highly advanced model with a smaller scale; and a large-scale model from an older generation. We begin by examining variations in responses across the four LLM families. We then, controlling for LLM family fixed effects, analyze how model generation and model scale influence patterns in LLM responses. As in Section 3.1, we conduct separate analyses for the six preference-based questions and the ten belief-based questions.

3.2.1. Heterogeneity across LLM families

We first examine variations in LLM response across the four LLM families. Fig. 2 provides preliminary graphical evidence of differences among the four benchmark models. For preference-based questions, Gemini 1.5 Pro, relative to GPT-4, produces a lower share of rational responses and a higher share of human-like responses. For belief-based questions, Llama 3 70B, relative to GPT-4, produces a lower share of rational responses and a higher share of human-like responses.

To formally assess heterogeneity in responses across the LLM families, we estimate a series of probit regressions using all twelve LLMs. The regression specification is:

$$\Pr(Y_{iqk} = 1) = \Phi(\alpha + \beta_1 \cdot Claude_i + \beta_2 \cdot Gemini_i + \beta_3 \cdot Llama_i + \epsilon_{iqk}) \tag{2}$$

for model i, question q, and iteration k, where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard Normal random variable. When studying how variation in LLM families affects the likelihood of observing a rational response, Y_{iqk}, the dependent variable in (2), is a binary variable that takes the value of one if model i's response to question q in iteration k is classified as rational, and zero otherwise. When studying how variation in LLM families affects the likelihood of observing a human-like response, Y_{iqk} is a binary variable that takes the value of one if model i's response to question q in iteration k is classified as human-like, and zero otherwise. For both cases, the independent variables—$Claude_i$, $Gemini_i$, and $Llama_i$—are indicators for the respective LLM families, with the LLM family of GPT serving as the omitted baseline.

[Place Table 4 about here]

Table 4 reports marginal effects from these regressions, where each reported coefficient represents the change in the predicted probability of observing $Y_{iqk} = 1$ that is associated with changing the LLM from GPT to Claude, Gemini, or Llama. Consistent with the heterogeneity observed from Fig. 2, for preference-based questions, Gemini models are 22.9% less likely to produce a rational response than GPT models; this effect is significant at the 1% level. At the same time, Gemini models are 16.7% more likely to produce a human-like response than GPT models; this effect is significant at the 5% level. Moreover, responses from Claude or Llama models to the preference-based questions are statistically similar to those from GPT models.

For belief-based questions, Llama models are 25.0% less likely to produce a rational response than GPT models; this effect is significant at the 5% level. Llama models are 21.0% more likely to produce a human-like response than GPT models; this effect is also significant at the 5% level. Finally, responses from Claude or Gemini models to the belief-based questions are statistically similar to those from GPT models. Overall, these findings highlight meaningful LLM family-level differences in responses to experimental questions drawn from cognitive psychology. In subsequent analyses of heterogeneity across model generations and scales, we control for LLM family fixed effects.

3.2.2. Heterogeneity across model generations and model scales

We next examine variations in LLM responses across model generations and model scales. Changes in model generation and scale capture key aspects of LLM development, including improvements in model architectures and advancements of reinforcement learning algorithms. To study the effect of model generation on LLM responses, we compare advanced models with older models of a similar scale. To study the effect of model scale, we compare large-scale models with smaller-scale ones of the same generation. In both analyses, we control for LLM family fixed effects.

[Place Fig. 3 about here]

We begin by presenting graphical evidence on differences in LLM responses across model generations and scales. Fig. 3 displays radar charts that summarize the number of preference-based questions and the number of belief-based questions for which each model produces predominantly rational or human-like responses. These visualizations offer a compact view of cross-model variation. For example, Claude 3 Haiku produces predominantly rational responses for three out of six preference-based questions, while Claude 3 Opus does not produce predominantly rational responses for any preference-based question.

The radar charts reveal a striking contrast between LLM responses to preference-based questions and their responses to belief-based questions. For preference-based questions, the left panel shows that, as LLMs become more advanced or larger, the number of questions with predominantly rational responses tends to decrease, while the number of questions with predominantly human-like responses increases. For belief-based questions, the right panel shows the *opposite* pattern: more advanced and larger-scale models tend to generate predominantly rational responses for a large number of questions.

To formally examine heterogeneity in LLM responses across model generations and scales, we estimate a series of probit regressions. In particular, we conduct two analyses. First, to study the effect of model generation on LLM responses, we restrict the sample to responses from either the four advanced large-scale models or the four older models. The regression specification is:

$$\Pr(Y_{iqk} = 1) = \Phi(\alpha + \beta \cdot Advanced_i + \gamma_f + \epsilon_{iqk}) \tag{3}$$

for model i, question q, and iteration k. When studying the effect of a change in model generation on the likelihood of observing a rational response, Y_{iqk} is a binary variable that takes the value of one if model i's response to question q in iteration k is classified as rational, and zero otherwise. When studying the effect of a change in model generation on the likelihood of observing a human-like response, Y_{iqk} is a binary variable that takes the value of one if model i's response to question q in iteration k is classified as human-like, and zero otherwise. For both cases, the key independent variable, $Advanced_i$, is an indicator for the advanced models. Moreover, γ_f captures LLM family fixed effects.

Second, to study the effect of model scale, we restrict the sample to responses from either the four advanced large-scale models or the four advanced smaller-scale models. The regression specification is:

$$\Pr(Y_{iqk} = 1) = \Phi(\alpha + \beta \cdot LargeScale_i + \gamma_f + \epsilon_{iqk}). \tag{4}$$

When studying the effect of a change in model scale on the likelihood of observing a rational response, Y_{iqk} is a binary variable that takes the value of one if model i's response to question q in iteration k is classified as rational, and zero otherwise. When studying the effect of a change in model scale on the likelihood of observing a human-like response, Y_{iqk} is a binary variable that takes the value of one if model i's response to question q in iteration k is classified as human-like, and zero otherwise. The key independent variable, $LargeScale_i$, is an indicator for the large-scale models.

[Place Table 5 about here]

Table 5 reports marginal effects from these regressions, where the reported coefficients represent the change in the predicted probability of observing $Y_{iqk} = 1$ that is associated with either moving from an older to an advanced model or from a smaller-scale to a large-scale model. The regression results are consistent with the variation in LLM responses observed in Fig. 3 across model generations and scales. For preference-based questions, Columns (1) to (4) in Panel A show that, as models become more advanced, their responses are less likely to be categorized as rational and more likely to be categorized as human-like; Columns (1) to (4) in Panel B shows that, as models become larger in scale, the same patterns occur—the models' responses are less likely to be rational and more likely to be human-like. For both Panels A and B, most coefficients reported in Columns (1) to (4) are statistically significant; the only exception is that, as models become larger, the increase in human-like responses to preference-based questions is insignificant.

For belief-based questions, Columns (5) to (8) in Panel A show that more advanced models generate responses that are more likely to be categorized as rational and less likely to be categorized as human-like; Columns (5) to (8) in Panel B shows the same patterns as models become larger in scale. For both Panels A and B, all coefficients reported in Columns (5) to (8) are statistically significant.

In summary, Fig. 3 and Table 5 show systematic heterogeneity in LLM responses across model generations and scales. As LLMs become more advanced or larger, their responses to preference-based questions become increasingly human-like, while their responses to belief-based questions become more rational. These opposing patterns highlight the importance of separately studying preferences and beliefs when evaluating LLM behavior.

3.3. LLM Responses to Questions from the Afrouzi et al. (2023) Experiments

We now examine LLM responses to questions from the three Afrouzi et al. (2023) experiments described in Section 2.1; we label these experiments as "Experiment 1," "Experiment 2," and

"Experiment 3." For each experiment, we simulate the autoregressive process:

$$x_t = \mu + \rho x_{t-1} + \epsilon_t$$

specified in (1), by setting μ, the constant term, to 0 and setting σ, the standard deviation of ϵ_t, to 20; these parameter values are adopted from Afrouzi et al. (2023). For ρ, the persistence parameter, we examine six values: 0, 0.2, 0.4, 0.6, 0.8, and 1. For each value, we generate 100 simulated paths, yielding a total of 600 paths per experiment.

For a given experiment and a given simulated path, we ask LLMs to make five rounds of forecasts. Take Experiment 1 as an example. In the first round, we present each LLM with a figure that displays the first 40 realizations of x_t from this simulated path, ranging from x_1 to x_{40}. Then, a prompt requests the LLM to provide its forecasts for the next two outcomes, x_{41} and x_{42}. The model's response is recorded to establish the beginning of a conversation history. In the second round, we first update the conversation history by adding the previous figure, prompt, and LLM response to it. We then present a new figure that extends the observed sequence to x_{41} and prompt the LLM to forecast the next two outcomes, x_{42} and x_{43}. We record the LLM's response and add it to the conversation history. This iterative process continues until five rounds of forecasts are completed.

To evaluate the extent to which the LLM's forecasts are biased, we estimate $\hat{\rho}$, the "perceived" autoregressive coefficient implied by these forecasts. Specifically, for each LLM i, each value of ρ, and each forecasting horizon s of 1, 2, or 5, we estimate the perceived persistence $\hat{\rho}$ using the following regression:

$$F_{it}x_{t+s} = c_{is} + (\hat{\rho}_{is})^s x_t + u_{is,t}, \tag{5}$$

where $F_{it}x_{t+s}$ represents model i's time-t forecast of x_{t+s}, x_t is the time-t realization of x, and $u_{is,t}$ is an error term.

Fig. 4 presents estimates based on LLM responses to questions from Experiment 1 of Afrouzi et al. (2023); here, we focus on short-term forecasts, with the forecasting horizon s set to one. The top panel displays the $\hat{\rho}$ values estimated for the three baseline models: GPT-4, Claude 3 Opus, and Gemini 1.5 Pro. The bottom panel displays the $\hat{\rho}$ values estimated for the three smaller-scale models: GPT-4o, Claude 3 Haiku, and Gemini 1.5 Flash.[20] For each $\hat{\rho}$ estimate, we also plot the 95% confidence interval. The results above are compared with a 45-degree line, which represents the persistence implied by full information rational expectations (FIRE).

[Place Fig. 4 about here]

Two observations emerge from Fig. 4. First, for the advanced large-scale models (top panel), LLM forecasts are largely rational: for each model and each ρ, the estimated $\hat{\rho}$ is close to the true ρ. Second, for the smaller-scale models (bottom panel), LLM forecasts are human-like: consistent with the findings of Afrouzi et al. (2023) for human participants, the persistence $\hat{\rho}$ implied by LLM forecasts is significantly higher than the true persistence ρ, and the difference between $\hat{\rho}$ and ρ is larger for lower values of ρ. The comparison between the top and bottom panels reinforces a pattern documented in Section 3.1 using cognitive psychology questions, namely that larger-scale models tend to generate more rational responses to belief-based questions.

We also examine LLM forecasts from Experiments 2 and 3. Fig. 5 plots $\hat{\rho}$ against ρ, with the top panel examining longer-term forecasts $F_{it}x_{t+5}$ from Experiment 2 and the bottom panel examining short-term forecasts $F_{it}x_{t+1}$ from Experiment 3, in which LLMs are given more detailed knowledge about the evolution of x_t.

[Place Fig. 5 about here]

The comparison between Fig. 4 and Fig. 5 yields two observations. First, for the three advanced large-scale models of GPT-4, Claude 3 Opus, and Gemini 1.5 Pro, LLMs' longer-term forecasts produce human-like biases that are absent in short-term forecasts: the top panel of Fig. 5 shows that, the persistence $\hat{\rho}$ implied by longer-term forecasts is significantly higher than the true persistence ρ, and the difference between $\hat{\rho}$ and ρ is larger for lower values of ρ. In comparison, as discussed before, the top panel of Fig. 4 shows that short-term forecasts are largely rational. Notice from Afrouzi et al. (2023) that human participants' longer-term forecasts are also more biased than their short-term forecasts.

Second, the comparison between the bottom panel of Fig. 5 and the top panel of Fig. 4 shows that providing detailed information about the data generating process can be counterproductive:

---

[20] Llama models do not support graphical inputs and are excluded from this analysis.

it gives rise to more human-like biases in LLM responses. Interestingly, this effect is unique to LLMs; Afrouzi et al. (2023) find that human forecasts are unaffected by the provision of additional information about the evolution of x_t. Section 4 provides more discussion of this finding.

3.4. LLM Responses to Questions from the Bose et al. (2022) Experiment

We next turn to Bose et al. (2022), who design an experiment in which human participants observe stock price trajectories and then make investment decisions. Specifically, the experiment generates price charts for 1,000 stocks randomly selected from the Center for Research in Security Prices (CRSP) database, using the stocks' daily returns from 2017. These price charts vary along two dimensions: the stock's cumulative annual return and its convexity score, which measures the curvature of the stock price trajectory. Each price chart is viewed by four human participants. Each participant is given an endowment of 1,000 monetary units and, after seeing a price trajectory, is asked: "How much of a 1,000 monetary unit endowment are you willing to invest in the stock for the following 12 months? The remainder will be put in a safe bank account."

Bose et al. (2022) construct two novel variables. The first is the "visual salience" of a stock price trajectory, denoted by "VS" and defined as a weighted average of the stock's past daily returns; the weights are computed using a machine learning algorithm that captures the visual attention paid to each return. The second is a related measure, $\mathbb{C}$orr(Returns, VS weights), defined as the correlation between past daily returns and their associated visual salience weights. The authors interpret both variables as reflecting non-traditional preferences—preferences that violate the Expected Utility framework—and show that these variables strongly predict participants' investment amounts, even after controlling for variables derived from alternative frameworks, including cumulative prospect theory (CPT; Tversky and Kahneman, 1992) and salience theory (Bordalo et al., 2012, 2013).

We replicate this setup with LLMs by presenting them with stock price trajectories and eliciting their investment amounts. Specifically, we follow the same procedure to select 1,000 stocks from CRSP and use their daily returns from 2017 to construct price charts; we apply the same machine learning algorithm to compute the two key variables, "VS" and $\mathbb{C}$orr(Returns, VS weights); and we construct the same control variables including those related to CPT (Tversky and Kahneman, 1992) and salience theory (Bordalo et al., 2012, 2013). Finally, as in the Afrouzi et al. (2023) experiments, we examine six of the twelve LLMs that support graphical input. Each LLM observes each of the

1,000 stock price charts four times, and in each instance, provides an investment amount, measured as the fraction of the endowed 1,000 monetary units that it is willing to invest in the stock.

[Place Table 6 about here]

Table 6 reports linear regression results that relate elicited investment amounts to a broad set of explantory variables for the advanced smaller-scale and advanced large-scale models of GPT, Claude, and Gemini. The key variables of interest are VS × $\mathbb{1}_{large}$ in Columns (1), (3), and (5) and $\mathbb{C}$orr(Returns, VS weights)×$\mathbb{1}_{large}$ in Columns (2), (4), and (6). The coefficients on these interaction terms are positive and mostly statistically significant.[21] These results indicate that, as an LLM increases in scale, its investment decisions depend more strongly on non-traditional preferences documented in Bose et al. (2022) for human behavior, reinforcing the pattern identified in Section 3.1 using cognitive psychology questions.[22]

4. Correcting LLM Biases

Section 3 documents that, for many preference-based and belief-based questions, LLM responses exhibit behavioral biases. In this section, we explore role-priming methods—instructing an LLM to view itself as a particular type of individual—as well as debiasing techniques that aim to correct these biases.

We begin by examining how role priming affects LLM responses. We consider two versions. The first instructs the LLMs to view themselves as rational investors; the second instructs them to view themselves as real-world retail investors. We implement each version by adding a single sentence at the beginning of the prompt. For the first version, the sentence is "When answering questions below, please think of yourself as a rational investor who makes decisions using the 'expected utility' framework." For the second version, the sentence is "When answering questions below, please think of yourself as a real-world retail investor who makes economic and financial decisions."

[Place Table 7 about here]

---

[21] For GPT, the coefficient on VS × $\mathbb{1}_{large}$ is positive but not statistically significant.

[22] The coefficients on VS and $\mathbb{C}$orr(Returns, VS weights) are negative. This pattern is consistent with smaller-scale models behaving rationally: in an equilibrium in which behavioral investors overweight stocks with high VS and $\mathbb{C}$orr(Returns, VS weights), these stocks become overvalued; as such, rational investors' allocations should load negatively on these variables.

Table 7 presents the effects of role priming on LLM behavior. Panel A reports the treatment effects of priming the LLMs to be rational investors. Averaged across the twelve LLMs, this role priming increases rational responses by 4.3% for preference-based questions (significant at the 5% level) and by 3.3% for belief-based questions (significant at the 10% level).[23] Panel B reports the treatment effects of priming the LLMs to be real-world retail investors. Averaged across the twelve LLMs, this role priming reduces rational responses by 3.9% for preference-based questions (significant at the 5% level) and has no statistically significant effect on responses to belief-based questions. Overall, Table 7 suggests that instructing LLMs to act as rational investors can reduce biases.

To better understand the mechanisms through which role priming affects LLM responses, we conduct a mediation analysis that examines whether role priming operates through intervening variables, referred to as mediators.[24] Specifically, we address three questions. First, we examine whether role priming affects the confidence levels that LLMs assign to their choices and their self-reported reasoning type, where type A corresponds to intuitive thinking and type B corresponds to analytical thinking and calculations. Second, we assess whether LLM confidence and reasoning type influence their responses. Third, we test whether including LLM confidence and reasoning type as control variables in regressions of LLM responses on role priming attenuates the estimated effect of role priming.

[Place Table 8 about here]

Table 8 presents the results. LLM confidence is measured by "high confidence," an indicator that equals one if an LLM assigns a confidence level greater than 0.9 to its choice, and reasoning type is measured by "system 2 thinking," an indicator that equals one if the LLM selects reasoning type B.[25] Panel A restricts the sample to LLM responses generated using either the baseline prompt or a treatment prompt that primes LLMs to be rational investors. Panel B restricts the sample to LLM responses generated using either the baseline prompt or a treatment prompt that primes LLMs to be real-world retail investors.

---

[23] We report treatment effects from regression specifications that include model fixed effects to account for differences across the twelve LLMs. The results are broadly similar without these controls.

[24] A classic example of mediation analysis is provided by Cohn et al. (2012).

[25] Our analysis includes all twelve LLMs but focuses on preference-based questions only, as Table 7 shows that the effects of role priming on LLMs' responses to belief-based questions are often insignificant.

Columns (1) and (2) show that role priming significantly affects LLM confidence and reasoning type. The rational-investor role-priming prompt increases "high confidence" by 6.6% and "system 2 thinking" by 11.4%; both effects are significant at the 1% level. In contrast, the retail-investor role-priming prompt deceases "high confidence" by 6.0% and "system 2 thinking" by 8.0%, with both effects significant at the 5% level. Columns (4) and (7) show that "high confidence" and "system 2 thinking" significantly influence LLM responses. For example, "system 2 thinking" increases rational responses by 40% in Panel A and by 38% in Panel B, while "high confidence" decreases human-like responses by 35% in Panel A and by 58% in Panel B. Finally, comparisons between Columns (3) and (5)—and between Columns (6) and (8)—show that including LLM confidence and reasoning type as control variables in regressions of LLM responses on role priming significantly attenuates the estimated effect of role priming. Taken together, Table 8 provides evidence that role priming affects LLM responses through confidence and reasoning type.

Table 9 explores two additional debiasing techniques. The first technique combines the sentence that primes LLMs to be rational investors with a detailed four-step procedure that guides the LLMs to choose a course of action rationally under the Expected Utility framework:

> "Please be reminded of the procedure of choosing a course of action under the 'expected utility' framework. For each course of action:
>
> (1) You list all possible wealth outcomes it could result; here, a wealth outcome accounts for existing wealth and any potential changes in wealth.
>
> (2) You compute the utility of each wealth outcome, using a globally concave utility function; note that the utility function focuses on total wealth outcomes rather than gains or losses alone.
>
> (3) You weigh the utility of each outcome by the probability of the outcome.
>
> (4) You sum up across outcomes to obtain the expected utility of the course of action.
>
> You repeat the four-step procedure above for each possible course of action and choose the course of action with the highest expected utility. When answering questions below, please provide the concrete steps you take for computing the expected utility of each course of action."

The second technique combines the sentence that primes LLMs to be rational investors with a summary of the key findings from Kahneman and Tversky (1979) that describe systematic human biases. The summary is generated by uploading the .pdf form of the original Kahneman and Tversky (1979) paper to an interactive GPT-4o chat box and then requesting a summary of the paper's key insights. The specific summary is given by:

> "Please be reminded of prospect theory, a framework that describes human decision-making. The main takeaway from Prospect Theory: An Analysis of Decision under Risk by Daniel Kahneman and Amos Tversky (1979) is that human decision-making under risk systematically deviates from the predictions of traditional expected utility theory. Instead of evaluating choices purely in terms of final wealth states, individuals evaluate gains and losses relative to a reference point.
>
> Key Insights:
>
> Certainty Effect – People overweight certain outcomes relative to probable ones, leading to risk aversion in gains and risk-seeking behavior in losses.
>
> Loss Aversion – Losses loom larger than equivalent gains, meaning the psychological impact of losing \$100 is greater than the pleasure of gaining \$100.
>
> Diminishing Sensitivity – The value function is concave for gains and convex for losses, meaning the impact of an additional dollar diminishes as amounts increase.
>
> Decision Weights vs. Probabilities – People do not evaluate probabilities linearly; they tend to overweight small probabilities (making lotteries attractive) and underweight moderate to high probabilities (explaining why they buy insurance).
>
> Isolation Effect – Decision-making is influenced by how choices are framed, leading to inconsistent preferences when identical problems are presented in different ways. This theory revolutionized behavioral economics by demonstrating that individuals do not always make rational choices based on maximizing expected utility but rather follow heuristics and biases shaped by psychological perceptions of risk and reward."

Importantly, as a debiasing technique, the goal of providing the key findings from Kahneman and Tversky (1979) is to help LLMs avoid making the same mistakes. Accordingly, we add the following sentence to the end of the above summary: "As a rational investor, you should avoid making the mistakes described in prospect theory."

[Place Table 9 about here]

Table 9 compares the baseline debiasing technique of simply priming LLMs to be rational investors with the two detailed debiasing techniques described above. The analysis in this table focuses only on the first three experimental questions listed in Table 1: these are prospect theory-related questions, one on diminishing sensitivity, one on loss aversion, and one on probability weighting.[26] Table 9 shows that the provision of the four-step procedure that guides LLMs to behave rationally is ineffective in reducing biases. Moreover, the provision of the key findings from Kahneman and Tversky (1979) *reduces* rational responses by about 29% and increases human-like responses by about 20%. Taken together, these results suggest that the provision of additional information—even if genuinely useful for decision making—does not always improve LLM performance: information overload can hinder an LLM's ability to produce rational responses. This is consistent with our finding in Section 3.3 that providing more information about the data generating process in the Afrouzi et al. (2023) experiments increases human-like biases in LLM responses.

In summary, simple role priming that instructs LLMs to act as rational investors is effective, although the magnitude of improvement is economically modest.[27] The two detailed debiasing techniques are either ineffective or counterproductive. Overall, our analysis in Section 4 suggests that debiasing LLMs remains a challenging task.

5. Conclusion

Artificial intelligence, especially generative AI epitomized by LLMs, has become increasingly important in social and economic activities. This paper calls for a structured study of the behavioral economics of AI. As a first step, we examine the behavior of four prominent families of LLMs—ChatGPT, Anthropic Claude, Google Gemini, and Meta Llama—by leveraging experimental designs from the cognitive psychology and experimental economics literatures.

We document systematic patterns in the behavioral biases that LLMs exhibit. For cognitive psychology and experimental economics questions that study human preferences, LLM responses

---

[26] We focus on these prospect theory-related questions for two reasons. First, LLM responses to these questions are often irrational, leaving room for debiasing. Second, one debiasing technique described above provides the key findings from prospect theory, which will mostly likely affect LLM responses to prospect theory-related questions.

[27] Table 3 shows that for preference-based questions, the baseline share of irrational (human-like) responses is often exceeds 50%, so a reduction of 3-4% reported in Table 7 is small relative to the baseline.

become increasingly irrational and human-like as models become more advanced or larger in scale. In contrast, for cognitive psychology and experimental economics questions that study human beliefs, the advanced large-scale LLMs generate responses that are largely rational. Moreover, we observe substantial heterogeneity in responses across the four LLM families.

We further explore role-priming and debiasing methods that may affect LLM behavior. A role-priming prompt that instructs LLMs to behave as rational investors who make decisions according to the Expected Utility framework is effective in reducing biases, while a prompt that instructs LLMs to behave as real-world retail investors leads to less rational responses. In both cases, role priming affects responses through changes in LLM confidence and reasoning type, although its overall impact is economically modest. Finally, providing genuinely useful information—either a detailed procedure that guides LLMs to choose a course of action rationally under the Expected Utility framework or a summary of key findings from Kahneman and Tversky (1979) that describe human biases—does not reduce LLM biases and can even be counterproductive.

References

Afrouzi, H., Kwon, S. Y., Landier, A., Ma, Y., Thesmar, D., 2023. Overreaction in Expectations: Evidence and Theory. Quarterly Journal of Economics 138, 1713–1764.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., et al, 2022. Constitutional AI: Harmlessness from AI Feedback. Working Paper.

Bail, C. A., 2024. Can Generative AI Improve Social Science? Proceedings of the National Academy of Sciences 121, e2314021121.

Bar-Hillel, M., 1979. The Role of Sample Size in Sample Evaluation. Organizational Behavior and Human Performance 24, 245–257.

Barberis, N., 2018. Psychology-Based Models of Asset Prices and Trading Volume. In: Bernheim, D., DellaVigna, S., Laibson, D. (Eds.), *Handbook of Behavioral Economics*, North Holland, Amsterdam, pp. 79–175.

Barberis, N., Thaler, R., 2003. A Survey of Behavioral Finance. In: Constantinides, G., Harris, M., Stulz, R. M. (Eds.), *Handbook of the Economics of Finance*, North Holland, Amsterdam, pp. 1053–1128.

Bauer, K., Liebich, L., Hinz, O., Kosfeld, M., 2023. Decoding GPT's Hidden 'Rationality' of Cooperation. Working Paper.

Binz, M., Schulz, E., 2023. Using Cognitive Psychology to Understand GPT-3. Proceedings of the National Academy of Sciences 120, e2218523120.

Bordalo, P., Gennaioli, N., Shleifer, A., 2012. Salience Theory of Choice Under Risk. Quarterly Journal of Economics 127, 1243–1285.

Bordalo, P., Gennaioli, N., Shleifer, A., 2013. Salience and Asset Prices. American Economic Review 103, 623–628.

Bose, D., Cordes, H., Nolte, S., Schneider, J., Camerer, C., 2022. Decision Weights for Experimental Asset Prices Based on Visual Salience. Review of Financial Studies 35, 5904–5126.

Bowen, D. E., Price, S. M., Stein, L. C., Yang, K., 2025. Measuring and Mitigating Racial Disparities in Large Language Model Mortgage Underwriting. Working Paper.

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, 25–37.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., et al, 2020. Language Models are Few-Shot Learners. Working Paper.

Bybee, L., 2025. The Ghost in the Machine: Generating Beliefs with Large Language Models. Working Paper.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., et al, 2024. A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology 15, 1–45.

Charness, G., Jabarian, B., List, J. A., 2023. Generation Next: Experimentation with AI. Working Paper.

Chen, S., Green, T. C., Gulen, H., Zhou, D., 2025. What Does ChatGPT Make of Historical Stock Returns? Extrapolation and Miscalibration in LLM Stock Return Forecasts. Working Paper.

Chen, Y., Kirshner, S., Ovchinnikov, A., Andiappan, M., Jenkin, T., 2024. A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? Working Paper.

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, e2316205120.

Choi, J. J., Laibson, D., Madrian, B. C., Metrick, A., 2004. For Better or for Worse: Default Effects and 401(k) Savings Behavior. In: Wise, D. A. (ed.), *Perspectives on the Economics of Aging*, University of Chicago Press, pp. 81–126.

Cohn, A., Engelmann, J., Fehr, E., Maréchal, M. A., 2012. Evidence for Countercyclical Risk Aversion: An Experiment with Financial Professionals. American Economic Review 105, 860–885.

Cook, T. R., Modig, Z., Kazinnik, S., Palmer, N., 2025. What Do LLMs Want? Working Paper.

Deaves, R., Lüders, E., Luo, G. Y., 2009. An Experimental Test of the Impact of Overconfidence and Gender on Trading Activity. Review of Finance 13, 555–575.

DellaVigna, S., Linos, E., 2022. RCTs to Scale: Comprehensive Evidence from Two Nudge Units. Econometrica 90, 81–116.

Ellsberg, D., 1961. Risk, Ambiguity, and the Savage Axioms. Quarterly Journal of Economics 75, 643–669.

Fan, C., Chen, J., Jin, Y., He, H., 2024. Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis. AAAI Conference on Artificial Intelligence 38, 17960–17967.

Frederick, S., Loewenstein, G., O'Donoghue, T., 2002. Time Discounting and Time Preference: A Critical Review. Journal of Economic Literature 40, 351–401.

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., Zarifhonarvar, A., 2025. Simulating the Survey of Professional Forecasters. Working Paper.

Kahneman, D., Tversky, A., 1973. On the Psychology of Prediction. Psychological Review 80, 237–251.

Kahneman, D., Tversky, A., 1979. Prospect Theory: An Analysis of Decision under Risk. Econometrica 47, 263–292.

Korinek, A., 2023. Generative AI for Economic Research: Use Cases and Implications for Economists. Journal of Economic Literature 61, 1281–1317.

Lian, C., Ma, Y., Wang, C., 2018. Low Interest Rates and Risk-Taking: Evidence from Individual Investment Decisions. Review of Financial Studies 32, 2107–2148.

Ma, D., Zhang, T., Saunders, M., 2023. Is ChatGPT Humanly Irrational? Working Paper.

Manning, B. S., Horton, J. J., 2025. General Social Agents. Working Paper.

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, e2313925121.

Moore, D. A., Healy, P. J., 2008. The Trouble with Overconfidence. Psychological Review 115, 502–517.

Ouyang, S., Yun, H., Zheng, X., 2024. How Ethical Should AI Be? How AI Alignment Shapes Risk Preferences of LLMs. Working Paper.

Rabin, M., 2002. Inference by Believers in the Law of Small Numbers. Quarterly Journal of Economics 117, 775–816.

Rapoport, A., Budescu, D. V., 1992. Generation of Random Series in Two-Person Strictly Competitive Games. Journal of Experimental Psychology: General 121, 352–363.

Rapoport, A., Budescu, D. V., 1997. Randomization in Individual Choice Behavior. Psychological Review 104, 603–617.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., et al, 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. Working Paper.

Shiffrin, R., Mitchell, M., 2023. Probing the Psychology of AI Models. Proceedings of the National Academy of Sciences 120, e2300963120.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., et al., 2020. Learning to Summarize from Human Feedback. Conference on Neural Information Processing Systems 34, 3008–3021.

Tegmark, M., 2017. Life 3.0: Being Human in the Age of Artificial Intelligence. Random House Audio Publishing Group.

Thaler, R. H., Sunstein, C. R., 2008. Nudge: Improving Decisions About Health, Wealth, and Happiness. Yale University Press.

Tomlinson, N., Laughridge, K., Dockar, B., 2024. Changing the Game: How AI is Poised to Transform Banking, Capital Markets. Wall Street Journal.

Tversky, A., Kahneman, D., 1974. Judgment under Uncertainty: Heuristics and Biases. Science 185, 1124–1131.

Tversky, A., Kahneman, D., 1981. The Framing of Decisions and the Psychology of Choice. Science 211, 453–458.

Tversky, A., Kahneman, D., 1983. Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. Psychological Review 90, 293–315.

Tversky, A., Kahneman, D., 1992. Advances in Prospect Theory: Cumulative Representation of Yncertainty. Journal of Risk and Uncertainty 5, 297–323.

Vafa, K., Rambachan, A., Mullainathan, S., 2024. Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function. Working Paper.

Vidal, N., 2023. How AI and LLMs are Streamlining Financial Services. Forbes.

Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press.

Wason, P. C., Johnson-Laird, P. N., 1972. Immediate Inferences with Quantifiers. In: Wason, P. C. (ed.), *Psychology of Reasoning: Structure and Content*, Harvard University Press, pp. 171–181.

Well, A. D., Pollatsek, A., Boyce, S. J., 1990. Understanding the Effects of Sample Size on the Variability of the Mean. Organizational Behavior and Human Decision Processes 47, 289–312.

Figures and Tables

Instructions:
Consider the following scenarios and respond according to the template provided. Please treat each scenario as completely separate from the other. The output should be a markdown code snippet formatted in the following schema, including the leading and trailing “`‘‘‘json`” and “`‘‘‘`” and should not include any note or comment:

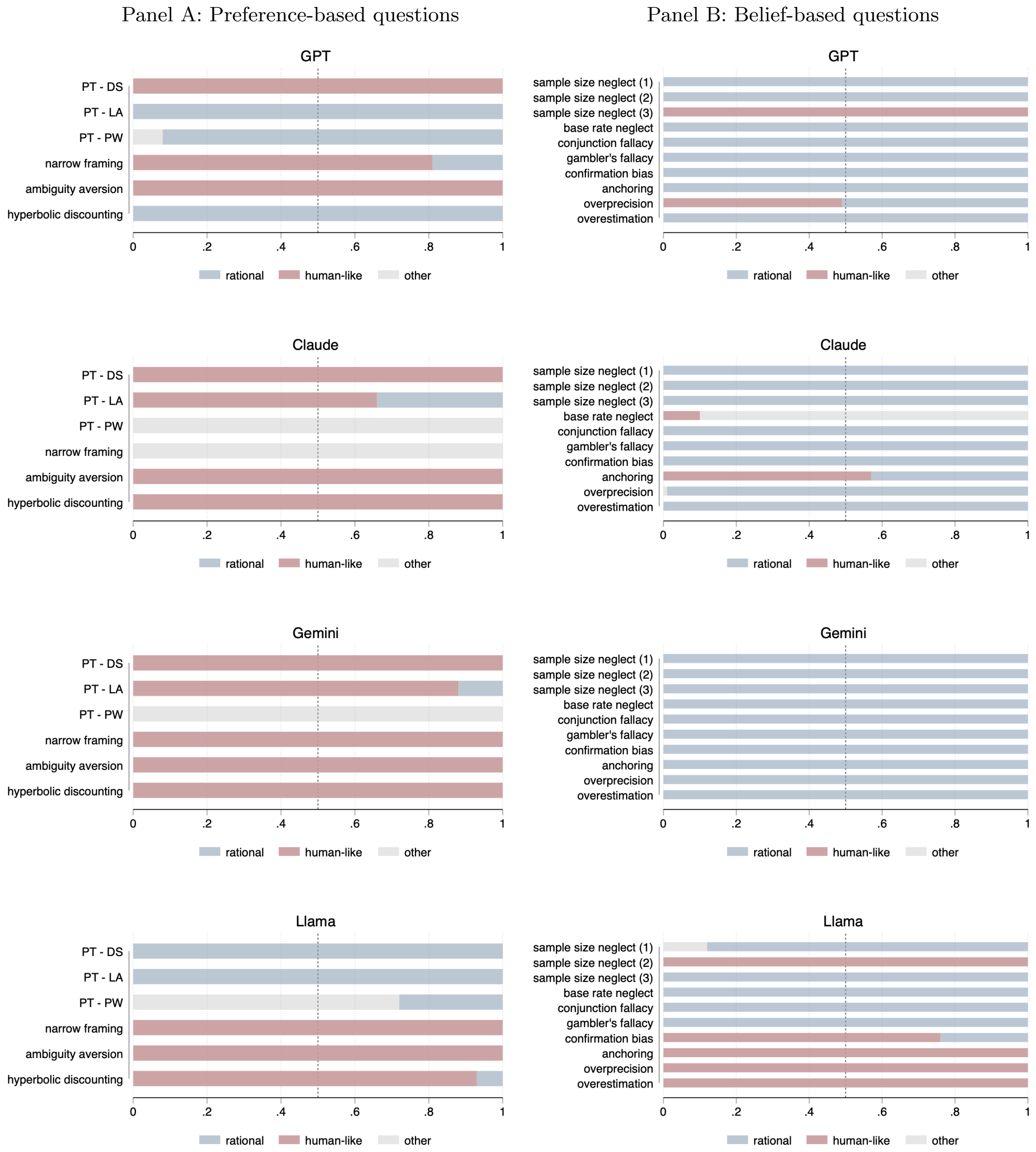
```
‘‘‘json
{
"Scenario A": {
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
},
"Scenario B": {
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
}
‘‘‘
```

Scenario A:
In addition to whatever you own, you have been given $1,000. You now need to choose between the following two options: option A ($1,000, 0.5), meaning winning $1,000 with 0.5 probability and winning zero with 0.5 probability, versus option B ($500), meaning winning $500 with certainty. Please answer as shown above. Indicate the choice you prefer (“A” or “B”), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type (“A” if your reasoning is based more on intuitive thinking, and “B” if your reasoning is based more on analytical thinking and calculations).

Scenario B:
Next, please consider a different scenario; please treat it as a completely separate scenario from the one you were just asked about. Specifically, please consider the following scenario. In addition to whatever you own, you have been given $2,000. You now need to choose between the following two options: option A (–$1,000, 0.5), meaning losing $1,000 with 0.5 probability and losing zero with 0.5 probability, versus option B: (–$500), meaning losing $500 with certainty. Please answer as shown above. Indicate the choice you prefer (“A” or “B”), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type (“A” if your reasoning is based more on intuitive thinking, and “B” if your reasoning is based more on analytical thinking and calculations).

Fig. 1. Example of prompt: Diminishing sensitivity of prospect theory.

This figure presents an example of a prompt that elicits LLM responses to a question designed by Kahneman and Tversky (1979) to document diminishing sensitivity, a key element of prospect theory.


Panel A: Preference-based questions

Panel B: Belief-based questions

Fig. 2. Proportion of LLM responses: Advanced large-scale models.

This figure plots the proportions of LLM responses categorized as rational (blue), human-like (red), or other (gray) for the four advanced large-scale LLMs: GPT-4, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3 70B. The left panel presents results for the six preference-based questions, while the right panel presents results for the ten belief-based questions.

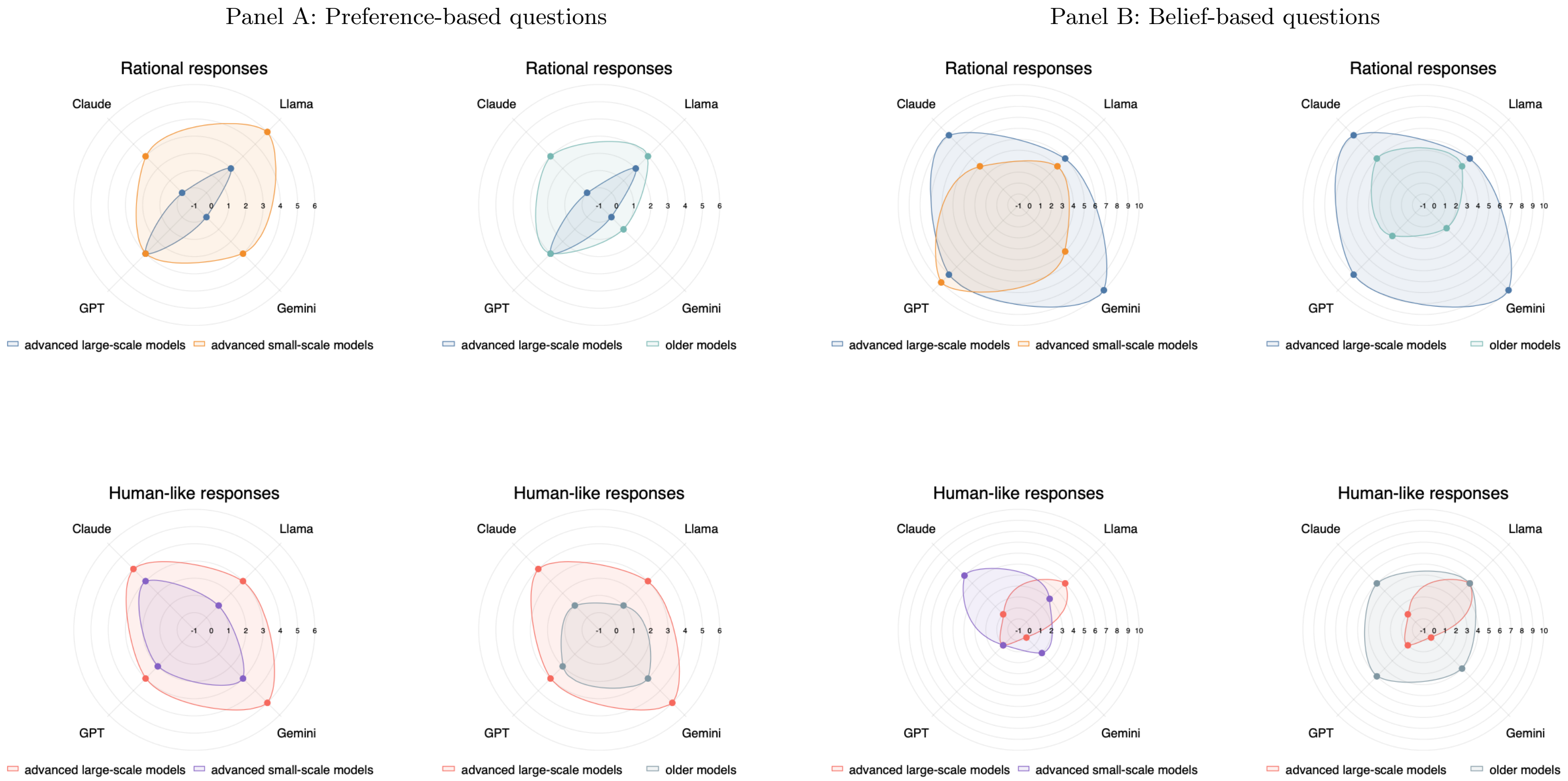


Fig. 3. Heterogeneity in LLM responses across model generations and model scales.

This figure presents radar charts that compare the number of questions for which each LLM produces predominantly rational responses (top row) or human-like responses (bottom row), separately for preference-based questions (left panel) and belief-based questions (right panel). Comparisons are made between advanced large-scale models and advanced smaller-scale models, and between advanced large-scale models and older models.

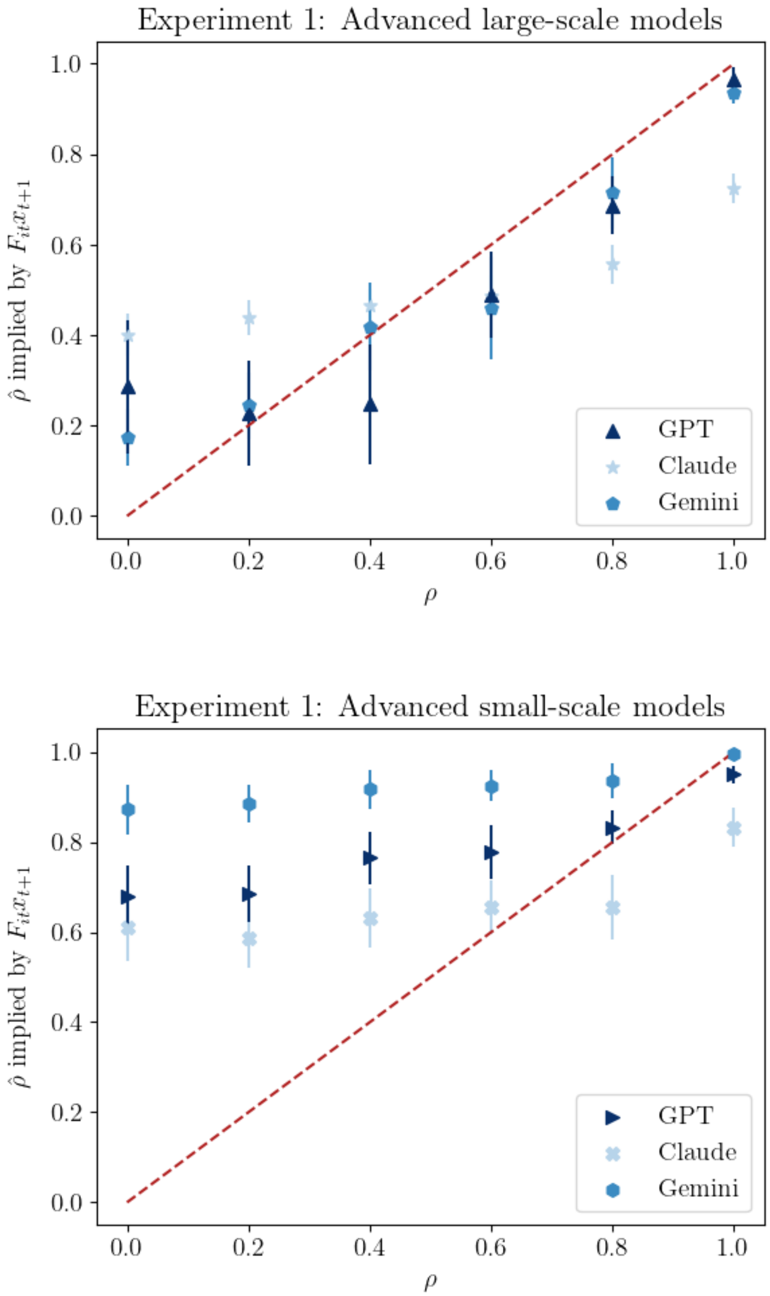


Fig. 4. LLM forecasts: Experiment 1 in Afrouzi et al. (2023).

The figure plots the perceived persistence $\hat{\rho}$ against the true ρ. Here, $\hat{\rho}$ is estimated using LLM forecasts collected from the setting of Experiment 1 in Afrouzi et al. (2023): the top panel reports results for the three advanced large-scale models of GPT-4, Claude 3 Opus, and Gemini 1.5 Pro, while the bottom panel reports results for the three advanced small-scale models of GPT-4o, Claude 3 Haiku, and Gemini 1.5 Flash. For each estimated $\hat{\rho}$, the vertical bar shows the 95% confidence interval. The procedure for estimating $\hat{\rho}$ is described in Section 3.3 of the main text. The red dashed line is a 45-degree line, which represents the persistence implied by full information rational expectations (FIRE).

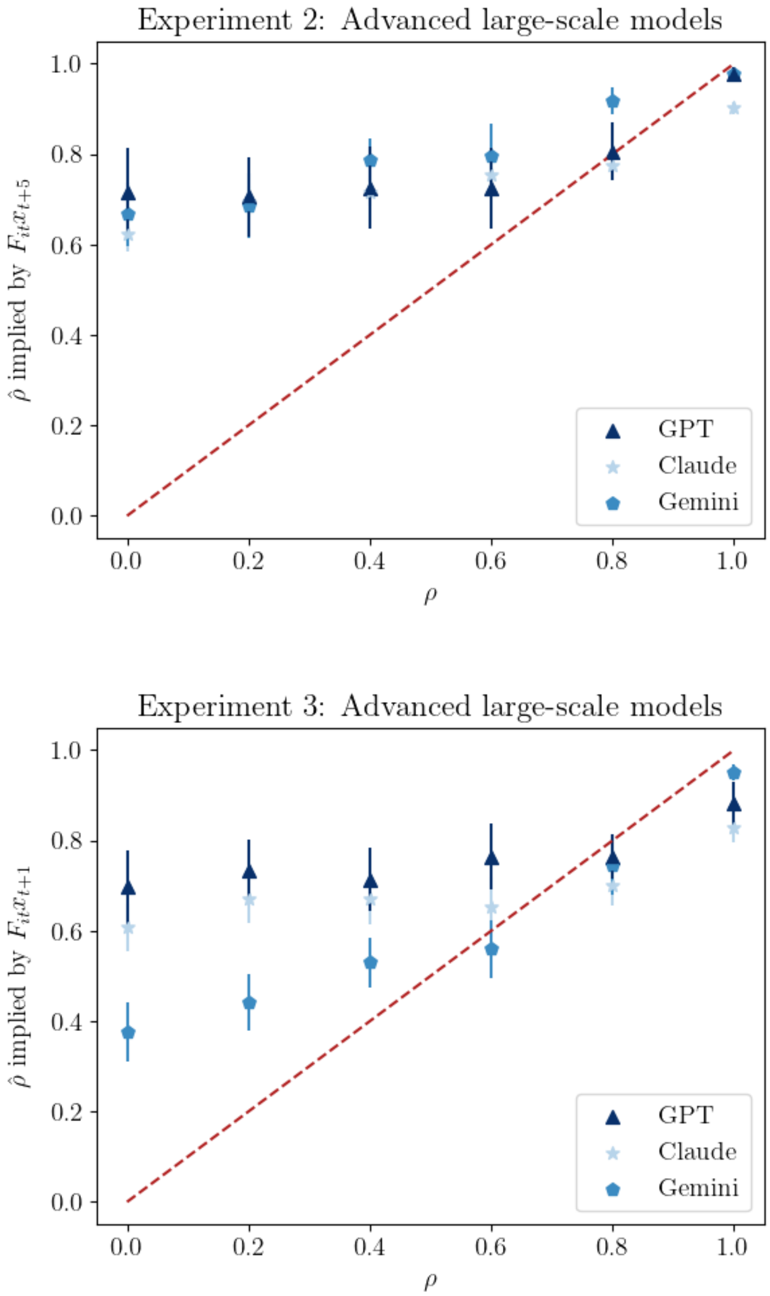


Fig. 5. LLM forecasts: Experiments 2 and 3 in Afrouzi et al. (2023).

The figure plots the perceived persistence $\hat{\rho}$ against the true ρ. Here, $\hat{\rho}$ is estimated using the LLM forecasts collected from the settings of Experiments 2 and 3 in Afrouzi et al. (2023): the top panel examines LLMs forecasts collected from the setting of Experiment 2, and the bottom panel examines LLM forecasts collected from the setting of Experiment 3. Both panels report results for the three advanced large-scale models of GPT-4, Claude 3 Opus, and Gemini 1.5 Pro. For each estimated $\hat{\rho}$, the vertical bar shows the 95% confidence interval. The procedure for estimating $\hat{\rho}$ is described in Section 3.3 of the main text. The red dashed line is a 45-degree line, which represents the persistence implied by full information rational expectations (FIRE).

Table 1. Summary of experimental questions from cognitive psychology.

This table summarizes the sixteen experimental questions drawn from the cognitive psychology literature. Specifically, Question 1 is based on Problems 11 and 12 in Kahneman and Tversky (1979) (page 273). Question 2 is based on an example of loss aversion discussed in Barberis and Thaler (2003) (page 1069). Question 3 is based on Problems 14 and 14′ in Kahneman and Tversky (1979) (page 281). Question 4 is a modified version of Problem 10 in Tversky and Kahneman (1981) (page 457). Question 5 is based on an example of hyperbolic discounting discussed in Frederick, Loewenstein, and O'Donoghue (2002) (page 361). Question 6 is based on Questions 3 and 4 of the Ellsberg (1961) experiment (pages 650 to 651). Question 7 is based on an experiment discussed in Tversky and Kahneman (1974) that documents sample size neglect as a form of representativeness heuristic (page 1125). Question 8 is based on Experiment 2 in Well, Pollatsek, and Boyce (1990) (page 297). Question 9 is based on Problem 10 in Bar-Hillel (1979) (page 255). Question 10 is based on an experiment designed in Kahneman and Tversky (1973) (page 241). Question 11 is based on an experiment designed in Tversky and Kahneman (1983) (pages 297 and 299). Question 12 is based on an experiment discussed in Rabin (2002) (page 781). Question 13 is based on a selection task discussed in Wason and Johnson-Laird (1972) (page 173). Question 14 is based on an experiment discussed in Tversky and Kahneman (1974) that documents anchoring (page 1128). Question 15 is based on a set of general knowledge questions adapted from Appendix C of Deaves, Lüders, and Luo (2009) (page 2). Finally, Question 16 follows the procedure discussed on pages 508 to 509 of Moore and Healy (2008) to document overestimation.

Panel A: Questions that study the psychology of preferences

Question number	Documented bias	Note
1	prospect theory - diminishing sensitivity	risk preferences
2	prospect theory - loss aversion	risk preferences
3	prospect theory - probability weighting	risk preferences
4	narrow framing	risk preferences
5	ambiguity aversion	risk preferences
6	hyperbolic discounting	time preferences

Panel B: Questions that study the psychology of beliefs

Question number	Documented bias
7	sample size neglect (1)
8	sample size neglect (2)
9	sample size neglect (3)
10	base rate neglect
11	conjunction fallacy
12	gambler's fallacy
13	confirmation bias
14	anchoring
15	overconfidence - overprecision
16	overconfidence - overestimation

Table 2. Description of large language models.

This table describes the twelve LLMs examined in our analysis. We group these models into four LLM families: ChatGPT, Anthropic Claude, Google Gemini, and Meta Llama. For each family, we treat the advanced large-scale models as baselines: GPT-4, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3 70B. We also analyze their smaller-scale versions—GPT-4o, Claude 3 Haiku, Gemini 1.5 Flash, and Llama 3 8B—as well as their predecessors—GPT-3.5 Turbo, Claude 2, Gemini 1.0 Pro, and Llama 2 70B. RLHF and RLAI denote "Reinforcement Learning from Human Feedback" and "Reinforcement Learning from AI," respectively. MMLU denotes "Massive Multitask Language Understanding" and it provides a benchmark score for evaluating LLM capabilities. Vision indicates whether a model supports graphical inputs.

Model	Release year	Number of parameters	Number of tokens	Instruction	Context window	MMLU	Vision
GPT-3.5 Turbo	2022	175 B	300 B	RLHF	16,385	70	No
GPT-4	2023	1T*	13T*	RLHF	128,000	86.5	Yes
GPT-4o	2024	-	13T*	RLHF	128,000	88.7	Yes
Claude 2	2023	200 B*	-	RLAI + RLHF	100,000	78.5	No
Claude 3 Opus	2024	1T*	-	RLAI + RLHF	200,000	86.8	Yes
Claude 3 Haiku	2024	20B*	-	RLAI + RLHF	200,000	75.2	Yes
Gemini 1.0 Pro	2024	100 B*	-	RLHF	32,000	-	Yes
Gemini 1.5 Pro	2024	1T*	-	RLHF	128,000	81.9	Yes
Gemini 1.5 Flash	2024	30 B*	-	RLHF	128,000	81.0	Yes
Llama 2 70B	2023	70 B	2 T	RLHF	4,096	68.9	No
Llama 3 70B	2024	70 B	15 T	RLHF	8,200	80.2	No
Llama 3 8B	2024	8 B	15 T	RLHF	8,200	68.4	No

*These numbers are unofficial and estimated.

Table 3. Rational responses versus human-like responses: Advanced large-scale models.

This table reports the proportion of responses classified as rational or human-like for the four advanced large-scale LLMs: GPT-4, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3 70B. Panel A presents results for the six preference-based questions, and Panel B presents results for the ten belief-based questions. The abbreviations "PT - DS," "PT - LA," and "PT - PW" denote "prospect theory - diminishing sensitivity," "prospect theory - loss aversion," and "prospect theory - probability weighting," respectively. Numbers in parentheses are p-values from a binomial test with the null hypothesis that the proportion of rational or human-like responses is less than or equal to 50%. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.1$.

Panel A: Preference-based questions

	GPT				Claude				Gemini				Llama			
	%rational		%human-like		%rational		%human-like		%rational		%human-like		%rational		%human-like	
PT - DS	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	1.00	(0.000)***	0.00	(1.000)
PT - LA	1.00	(0.000)***	0.00	(1.000)	0.34	(1.000)	0.66	(0.001)***	0.12	(1.000)	0.88	(0.000)***	1.00	(0.000)***	0.00	(1.000)
PT - PW	0.92	(0.000)***	0.00	(1.000)	0.00	(1.000)	0.00	(1.000)	0.00	(1.000)	0.00	(1.000)	0.28	(1.000)	0.00	(1.000)
narrow framing	0.19	(1.000)	0.81	(0.000)***	0.00	(1.000)	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***
ambiguity aversion	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***
hyperbolic discounting	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.07	(1.000)	0.93	(0.000)***

Panel B: Belief-based questions

	GPT				Claude				Gemini				Llama			
	%rational		%human-like		%rational		%human-like		%rational		%human-like		%rational		%human-like	
sample size neglect (1)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	0.88	(0.000)***	0.00	(1.000)
sample size neglect (2)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***
sample size neglect (3)	0.00	(1.000)	1.00	(0.000)***	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)
base rate neglect	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	0.10	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)
conjunction fallacy	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)
gambler's fallacy	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)
confirmation bias	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	0.24	(1.000)	0.76	(0.000)***
anchoring	1.00	(0.000)***	0.00	(1.000)	0.43	(0.933)	0.57	(0.097)*	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***
overprecision	0.51	(0.460)	0.49	(0.618)	0.99	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***
overestimation	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***

Table 4. Heterogeneity in responses across LLM families.

This table reports marginal effects from the probit regressions:

$$\Pr(Y_{iqk} = 1) = \Phi(\alpha + \beta_1 \cdot Claude_i + \beta_2 \cdot Gemini_i + \beta_3 \cdot Llama_i + \epsilon_{iqk})$$

for model i, question q, and iteration k, where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard Normal random variable. For Columns (1) and (3), Y_{iqk} is a binary variable that takes the value of one if model i's response to question q in iteration k is classified as rational, and zero otherwise. For Columns (2) and (4), Y_{iqk} is a binary variable that takes the value of one if model i's response to question q in iteration k is classified as human-like, and zero otherwise. For both cases, the independent variables—$Claude_i$, $Gemini_i$, and $Llama_i$—are indicators for the respective LLM families, with the LLM family of GPT serving as the omitted baseline category. The reported coefficients represent the change in the predicted probability of observing $Y_{iqk} = 1$ that is associated with changing the LLM from GPT to Claude, Gemini, or Llama. Standard errors, clustered at the question level, are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.1$.

	(1)	(2)	(3)	(4)
Dep. var:	LLM response is characterized as			
	Rational	Human-like	Rational	Human-like
Sample:	Preference-based questions		Belief-based questions	
Claude	–0.126	–0.0483	–0.0997	0.126
	(0.083)	(0.118)	(0.084)	(0.102)
Gemini	–0.229***	0.167**	–0.0800	0.0107
	(0.065)	(0.077)	(0.049)	(0.051)
Llama	0.0816	–0.141	–0.250**	0.210**
	(0.150)	(0.127)	(0.098)	(0.088)
Baseline LLM family:	GPT			
Observations	7,150	7,150	12,000	12,000
Pseudo R-squared	0.043	0.037	0.025	0.026

Table 5. Heterogeneity in responses across model generations and model scales.

This table reports marginal effects from the probit regressions specified in equations (3) and (4) in the main text. For Columns (1), (2), (5), and (6), the dependent variable is a binary variable that takes the value of one if an LLM response is classified as rational, and zero otherwise. For Columns (3), (4), (7), and (8), the dependent variable is a binary variable that takes the value of one if an LLM response is classified as human-like, and zero other wise. Columns (1) to (4) report results for preference-based questions and Columns (5) to (8) report results for belief-based questions. Panel A compares advanced large-scale models with older models. In this case, the sample is restricted to LLM responses from either the advanced large-scale models or the older models, and the key independent variable is an indicator for the advanced models, with older models serving as the baseline. Panel B compares large-scale models with smaller ones. In this case, the sample is restricted to LLM responses from either the advanced large-scale models or the advanced smaller-scale models, and the key independent variable is an indicator for the large-scale models. Standard errors, clustered at the question level, are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.1$.

Panel A: Advanced models versus older models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var:	LLM response is characterized as							
	Rational	Rational	Human-like	Human-like	Rational	Rational	Human-like	Human-like
Sample:	Preference-based questions				Belief-based questions			
Advanced	–0.223*	–0.231**	0.272**	0.273**	0.407***	0.409***	–0.327***	–0.333***
	(0.121)	(0.116)	(0.127)	(0.126)	(0.127)	(0.125)	(0.104)	(0.102)
LLM family FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4,800	4,800	4,800	4,800	8,000	8,000	8,000	8,000
Pseudo *R*-squared	0.042	0.120	0.055	0.107	0.133	0.162	0.097	0.134

Panel B: Large-scale models versus smaller-scale models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var:	LLM response is characterized as							
	Rational	Rational	Human-like	Human-like	Rational	Rational	Human-like	Human-like
Sample:	Preference-based questions				Belief-based questions			
Large	–0.321***	–0.331***	0.212	0.216	0.240***	0.239***	–0.155**	–0.157**
	(0.093)	(0.091)	(0.130)	(0.132)	(0.092)	(0.090)	(0.074)	(0.073)
LLM family FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4,750	4,750	4,750	4,750	8,000	8,000	8,000	8,000
Pseudo *R*-squared	0.081	0.153	0.033	0.066	0.054	0.144	0.029	0.117

Table 6. LLM investment decisions: Experiment in Bose et al. (2022).

This table reports linear regression results that relate investment amounts elicited from LLMs—measured as the fraction of the endowed 1,000 monetary units that the LLM is willing to invest in a stock—to a broad set of explanatory variables. The key variables are VS, $\mathbb{C}$orr(Returns, VS weights), and their interaction with $\mathbb{1}_{large}$, where VS is a weighted average of the stock's past daily returns, with weights computed using the SAM algorithm described in Bose et al. (2022) that captures the visual attention paid to each return; $\mathbb{C}$orr(Returns, VS weights) is the correlation between past daily returns and their associated visual salience weights; and $\mathbb{1}_{large}$ is an indicator for the large-scale models. All control variables are constructed following the procedure in Bose et al. (2022). Robust standard errors are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.1$.

Dep. var:	Investment amount (%)					
	GPT		Claude		Gemini	
	(1)	(2)	(3)	(4)	(5)	(6)
VS	-0.0105		-0.0002		-0.0301^{***}	
	(0.0065)		(0.0048)		(0.0052)	
CPT	0.0841^{***}		0.0218^{***}		0.0780^{***}	
	(0.0081)		(0.0060)		(0.0065)	
Salience	0.0017		-0.0037		-0.0249^{***}	
	(0.0053)		(0.0043)		(0.0048)	
VS × $\mathbb{1}_{large}$	0.0089		0.0173^{**}		0.0225^{**}	
	(0.0073)		(0.0081)		(0.0099)	
CPT × $\mathbb{1}_{large}$	-0.0567^{***}		-0.0184^{*}		-0.0039	
	(0.0092)		(0.0103)		(0.0123)	
Salience × $\mathbb{1}_{large}$	-0.0098		0.0131^{*}		0.0166^{*}	
	(0.0061)		(0.0074)		(0.0087)	
Corr(Returns, VS weights)		-0.0315^{***}		-0.0072^{***}		-0.0280^{***}
		(0.0023)		(0.0016)		(0.0021)
Corr(Returns, CPT weights)		0.0255^{***}		0.0091^{***}		0.0182^{***}
		(0.0044)		(0.0033)		(0.0039)
Corr(Returns, Salience weights)		0.0206^{***}		-0.0007		-0.0008
		(0.0039)		(0.0031)		(0.0037)
Corr(Returns, VS weights) × $\mathbb{1}_{large}$		0.0260^{***}		0.0082^{***}		0.0099^{***}
		(0.0025)		(0.0031)		(0.0030)
Corr(Returns, CPT weights) × $\mathbb{1}_{large}$		-0.0085		-0.0029		0.0224^{***}
		(0.0052)		(0.0066)		(0.0061)
Corr(Returns, Salience weights) × $\mathbb{1}_{large}$		-0.0273^{***}		0.0069		-0.0002
		(0.0046)		(0.0060)		(0.0059)
Constant	0.5865^{***}	0.5865^{***}	0.4806^{***}	0.4806^{***}	0.5990^{***}	0.5990^{***}
	(0.0015)	(0.0015)	(0.0013)	(0.0013)	(0.0014)	(0.0014)
$\mathbb{1}_{large}$	-0.0908^{***}	-0.0908^{***}	-0.1568^{***}	-0.1568^{***}	0.0710^{***}	0.0710^{***}
	(0.0018)	(0.0017)	(0.0024)	(0.0024)	(0.0022)	(0.0022)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,000	8,000	8,000	8,000	8,000	8,000
Adjusted *R*-squared	0.717	0.722	0.458	0.457	0.692	0.695

Table 7. Treatment effects of role-priming prompts.

This table reports marginal effects from a series of probit regressions. For Columns (1), (2), (5), and (6), the dependent variable is a binary variable that takes the value of one if an LLM response is classified as rational, and zero otherwise; for Columns (3), (4), (7), and (8), the dependent variable is a binary variable that takes the value of one if an LLM response is classified as human-like, and zero otherwise. Regressions in Columns (1) to (4) are for preference-based questions, and regressions in Columns (5) to (8) are for belief-based questions; each regression uses responses from all twelve LLMs. Panel A restricts the sample to LLM responses generated using either the baseline prompt or a treatment prompt that primes LLMs to be rational investors. Panel B restricts the sample to LLM responses generated using either the baseline prompt or a treatment prompt that primes LLMs to be real-world retail investors. For both panels, the key independent variable is an indicator for the treatment prompt, with the baseline prompt serving as the omitted baseline category. Standard errors, clustered at the question level, are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.1$.

Panel A: Role-priming prompt (rational investor)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var:	LLM response is characterized as							
	Rational	Rational	Human-like	Human-like	Rational	Rational	Human-like	Human-like
Sample:	Preference-based questions				Belief-based questions			
Role-priming prompt	0.0439***	0.0430**	–0.0418*	–0.0405*	0.0331*	0.0325*	–0.0087	–0.0067
	(0.017)	(0.017)	(0.021)	(0.021)	(0.019)	(0.019)	(0.025)	(0.024)
Model FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	14,308	14,308	14,308	14,308	23,993	23,993	23,993	23,993
Pseudo R-squared	0.001	0.155	0.001	0.098	0.001	0.184	0.000	0.153

Panel B: Role-priming prompt (retail investor)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var:	LLM response is characterized as							
	Rational	Rational	Human-like	Human-like	Rational	Rational	Human-like	Human-like
Sample:	Preference-based questions				Belief-based questions			
Role-priming prompt	–0.0361*	–0.0388**	0.0150	0.0152	0.0010	–0.0021	0.0052	0.0084
	(0.019)	(0.019)	(0.024)	(0.025)	(0.018)	(0.018)	(0.020)	(0.020)
Model FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	14,310	14,310	14,310	14,310	23,999	23,999	23,999	23,999
Pseudo R-squared	0.001	0.165	0.000	0.101	0.000	0.163	0.000	0.143

Table 8. Treatment effects of role-priming prompts: Mediation analysis.

This table reports marginal effects from a series of probit regressions that analyze the mechanisms through which role priming affects LLM responses. For Column (1), the dependent variable is an indicator for "high confidence," which equals one if an LLM assigns a confidence level greater than 0.9 to its choice (the median confidence level in our sample). For Column (2), the dependent variable is an indicator for "system 2 thinking," which equals one if the LLM selects reasoning type B. For Columns (3) to (5), the dependent variable is a binary variable that takes the value of one if an LLM response is classified as rational, and zero otherwise; for Columns (6) to (8), the dependent variable is a binary variable that takes the value of one if an LLM response is classified as human-like, and zero otherwise. All regressions include responses from all twelve LLMs but focus on preference-based questions only. Panel A restricts the sample to LLM responses generated using either the baseline prompt or a treatment prompt that primes LLMs to be rational investors. Panel B restricts the sample to LLM responses generated using either the baseline prompt or a treatment prompt that primes LLMs to be real-world retail investors. For both panels, key independent variables include an indicator for the treatment prompt (with the baseline prompt serving as the omitted category), a "high confidence" indicator, and a "system 2 thinking" indicator. Standard errors, clustered at the question level, are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.1$.

Panel A: Role-priming prompt (rational investor)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var:	High confidence	System 2 thinking	LLM response is characterized as					
			Rational	Rational	Rational	Human-like	Human-like	Human-like
Sample:	Preference-based questions							
Role-priming prompt	0.0656***	0.114***	0.0430**		0.000718	–0.0405*		0.00417
	(0.020)	(0.041)	(0.017)		(0.010)	(0.021)		(0.015)
High confidence				0.143	0.143		–0.350*	–0.351*
				(0.142)	(0.141)		(0.187)	(0.187)
System 2 thinking				0.400***	0.400***		–0.201*	–0.202*
				(0.124)	(0.124)		(0.105)	(0.104)
Model FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	11,911	13,108	14,308	14,307	14,307	14,308	14,307	14,307
Pseudo R-squared	0.166	0.278	0.155	0.217	0.217	0.098	0.165	0.165

Panel B: Role-priming prompt (retail investor)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var:	High confidence	System 2 thinking	LLM response is characterized as					
			Rational	Rational	Rational	Human-like	Human-like	Human-like
Sample:	Preference-based questions							
Role-priming prompt	–0.0601**	–0.0797**	–0.0388**		–0.00529	0.0152		–0.0182
	(0.031)	(0.033)	(0.019)		(0.025)	(0.025)		(0.034)
High confidence				0.322***	0.321**		–0.581***	–0.583***
				(0.123)	(0.125)		(0.180)	(0.182)
System 2 thinking				0.377**	0.376**		–0.199	–0.202
				(0.156)	(0.156)		(0.129)	(0.129)
Model FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,728	14,310	14,310	14,310	14,310	14,310	14,310	14,310
Pseudo R-squared	0.135	0.301	0.165	0.256	0.256	0.101	0.199	0.199

Table 9. Comparison of debiasing techniques: Prospect theory-related questions.

This table reports marginal effects from a series of probit regressions. For Columns (1) and (2), the dependent variable is a binary variable that takes the value of one if an LLM response is classified as rational, and zero otherwise; for Columns (3) and (4), the dependent variable is a binary variable that takes the value of one if an LLM response is classified as human-like, and zero otherwise. Regressions are estimated using LLM responses to prospect theory-related questions only; each regression uses responses from all twelve LLMs. Panel A restricts the sample to LLM responses generated using either the baseline prompt (omitted category) or a treatment prompt that primes LLMs to be rational investors. Panel B restricts the sample to LLM responses generated using either the rational-investor role-priming prompt (omitted category) or an instruction-based prompt that combines the sentence that primes LLMs to be rational investors with a detailed four-step procedure that guides LLMs to choose actions rationally. Panel C restricts the sample to the LLM responses generated using either the rational-investor role-priming prompt (omitted category) or a knowledge-enrichment prompt that combines the sentence that primes LLMs to be rational investors with a summary of the key findings from Kahneman and Tversky (1979) that describes systematic human biases. Standard errors, clustered at the question level, are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.1$.

Panel A: Role-priming prompt (rational investor)

	(1)	(2)	(3)	(4)
Dep. var:	LLM response is characterized as			
	Rational	Rational	Human-like	Human-like
Sample:	Prospect theory-related questions			
Role-priming prompt	0.0375***	0.0401***	–0.0225**	–0.0267**
	(0.007)	(0.007)	(0.011)	(0.012)
Model FE	No	Yes	No	Yes
Observations	7,195	7,195	7,195	6,595
Pseudo *R*-squared	0.001	0.231	0.001	0.150

Panel B: Instruction-based prompt

	(1)	(2)	(3)	(4)
Dep. var:	LLM response is characterized as			
	Rational	Rational	Human-like	Human-like
Sample:	Prospect theory-related questions			
Instruction-based prompt	–0.0987	–0.0985	0.0836	0.0874
	(0.078)	(0.074)	(0.085)	(0.087)
Model FE	No	Yes	No	Yes
Observations	7,195	7,195	7,195	6,595
Pseudo *R*-squared	0.007	0.241	0.008	0.211

Panel C: Knowledge-enrichment prompt

	(1)	(2)	(3)	(4)
Dep. var:	LLM response is characterized as			
	Rational	Rational	Human-like	Human-like
Sample:	Prospect theory-related questions			
Knowledge-enrichment prompt	–0.291***	–0.287***	0.203*	0.204*
	(0.055)	(0.052)	(0.112)	(0.106)
Model FE	No	Yes	No	Yes
Observations	7,191	7,191	7,191	7,191
Pseudo *R*-squared	0.070	0.262	0.038	0.149