# Behavioral Economics of AI: LLM Biases and Corrections*

Pietro Bini[a] Lin William Cong[a,b] Xing Huang[c] Lawrence J. Jin[a,b]

[a] *SC Johnson College of Business, Cornell University*
[b] *National Bureau of Economic Research*
[c] *Olin Business School, Washington University in Saint Louis*

First draft: January 2025; this draft: April 2025.

**ABSTRACT**

Do generative AI models, as epitomized and popularized by large language models (LLMs), exhibit systematic behavioral biases in economic and financial decisions? If so, how can we mitigate these biases? Following the cognitive psychology literature and the experimental economics studies, we conduct the most comprehensive set of experiments to date—originally designed to document human biases—on prominent LLM families with variations in model version and scale. We document systematic patterns in the behavioral biases that LLMs exhibit. For experiments concerning the psychology of preferences, LLM responses become increasingly irrational and human-like as the models become more advanced or larger; however, for experiments concerning the psychology of beliefs, the most advanced large-scale models frequently generate rational responses. Further exploring various methods for correcting these behavioral biases reveals that prompting LLMs to make rational decisions according to the Expected Utility framework seems the most effective.

*JEL classification*: D03, G02, G11, G41
*Keywords*: AI, Behavioral Biases, Beliefs, Preferences, LLMs

*We are grateful to Nicholas Barberis, James Choi, William Goetzmann, Siew Hong Teoh, Luyao Zhang, and seminar participants at the University of California Los Angeles, the Conference for Financial Economics and Accounting, and the ADEFT-XueShuo Winter Institute in AI for Social Sciences & the Economics of AI 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.

# 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 and management (Vidal, 2023; Tomlinson, Laughridge, and Dockar, 2024). Researchers are investigating the potential for LLMs to enhance experimentation that studies human behavior (Charness, Jabarian, and List, 2023; Korinek, 2023; Bail, 2024). However, little is known about how AI algorithms and agents behave systematically, especially in economic and financial decisions, let alone whether their behavior closely resembles that of humans. Understanding the "behavioral economics" of AI—potentially a new intelligent life form (Tegmark, 2017)—starting with LLMs is urgent and crucial for assessing and improving the technology's utility, safety, and appropriateness.

Recent studies have started to examine the reliability of LLMs in decision making, with a focus on specific behavior of ChatGPT.[1] Our paper not only adds to these studies but also conceptually introduce a new field—behavioral economics of AI—through establishing its 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 families of LLMs; we systematically compare LLM responses with both rational responses and human responses; and we explore methods for correcting their biases. An important goal of the paper is to develop a public database of experimental questions for ongoing evaluations of behavioral biases in various LLMs.

We begin by exploring two broad approaches for conducting experiments that allow us to document the behavioral biases of LLMs. First, we draw on the cognitive psychology literature, originated by Ellsberg (1961) and Kahneman and Tversky (1973, 1979), that uses carefully designed experimental questions to assess the 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. And our choice of questions ensures the inclusion of those used to document the psychological biases that are first-order important

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

in financial markets.[2] For each question, we design a prompt that is applicable to LLMs, hence allowing us to elicit responses from these models and analyze their behavior. Next, we turn to recent experimental economics studies, which, compared to the cognitive psychology literature, include 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 at 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 of LLMs, we consider two variations. First, we examine an advanced version of the model alongside an older version; this allows us to study the time-series variation in the model's degree of behavioral biases. Second, for the advanced model, we compare one version with a large parameter scale to another with a smaller scale; this allows us to study the cross-sectional variation in the model's degree of behavioral biases.

Analyzing the responses from the LLMs gives rise to five observations. First, when asked questions from the cognitive psychology literature that document biases in preferences, the LLMs' answers exhibit a clear pattern: as we progress to more advanced models or those with a larger scale, the 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 but 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, when asked questions from the cognitive psychology literature that document biases in beliefs, the LLMs' answers exhibit the *opposite* pattern: for more advanced models or those with a larger scale, the responses are increasingly rational. For example, Gemini 1.5 Pro, a highly advanced large-scale LLM, answers ten out of ten belief-based questions correctly. In comparison, Gemini 1.5 Flash, another advanced but smaller model, answers five out of the ten questions correctly, while Gemini 1.0 Pro, an older version, answers only two out of the ten questions correctly. Overall, three out of the four advanced large-scale LLMs we examine—GPT-4, Claude 3 Opus, and Gemini

[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] They are also the extant LLMs when we started our study in 2023.

1.5 Pro—produce by and large rational answers to belief-based questions.

Third, we observe substantial heterogeneity in LLM responses when comparing responses across the four different families of LLMs. For preference-based questions, the responses from Gemini are less rational and more human-like compared to those from ChatGPT, while the responses from Claude or Llama are by and large similar to those from ChatGPT. For belief-based questions, the responses from Meta Llama 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 the LLMs' responses to questions from experimental economics that document biases in investor beliefs. Specifically, we follow the recent work of Afrouzi et al. (2023) by asking the LLMs to first observe a sequence of past realizations of a random variable and then forecast its future realizations; the time-series evolution of this random variable is governed by an autoregressive process. We show that, for the advanced small-scale LLMs—GPT-4o, Claude 3 Haiku, and Gemini 1.5 Flash—their forecasts are irrational and human-like: they perceive an autoregressive process that is more persistent than the true process. Interestingly, compared to these small-scale LLMs, the larger-scale models—GPT-4, Claude 3 Opus, and Gemini 1.5 Pro—generate forecasts that are more rational: their perceived persistence of the autoregressive process is similar to the true persistence.[4] This finding suggests that, for belief-based questions from *both* the cognitive psychology literature and the experimental economics studies, the LLMs' responses become more rational as we progress from small-scale models to larger-scale ones.

Finally, we explore methods for correcting the observed behavioral biases. Among the methods, one seems effective while the others are not. The effective method involves a brief role-priming instruction that asks a LLM to think of itself as a rational investor who makes decisions using the Expected Utility framework; such an instruction is provided prior to the LLM answering any question. We find that, relative to the baseline results, adding such a role-priming instruction makes the LLM responses more rational and less human-like, for both the preference-based questions and the belief-based questions. The other methods involve combining the brief sentence that primes a LLM to be a rational investor with the provision of additional bias-reducing information. These

[4]For the questions designed in Afrouzi et al. (2023), eliciting responses from LLMs requires providing the models with graphical inputs—figures that display a sequence of past realizations of a random variable. Currently, six out of the twelve LLMs we examine do not support graphical inputs. As such, we do not run the questions on these LLMs. See Section 2.3 for a detailed discussion.

methods are found to be ineffective in reducing biases, suggesting that information overload might hinder a LLM's ability to give rational responses.

The five observations stated above are descriptive, albeit informative. Although a full discussion of the underlying mechanisms that drive our 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 in part due to the fact that advanced and large-scale LLMs are increasingly based 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 in part due to the fact that the larger training data and the greater computational power of advanced and large-scale LLMs enable these models to better identify ground truth in statistics on which they base their responses to belief-based questions. Studying these conjectures may inform future LLM designs.

**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. Correspondingly, a strand of research aims at developing and understanding methods that debias human participants (Choi et al., 2004; Thaler and Sunstein, 2008; DellaVigna and Linos, 2022). We expand the boundaries of these research fields by moving beyond understanding human behavior to study a new field of behavioral economics of AI. Doing so helps address two fundamental issues: $(i)$ the rapid advancement of GenAI has led researchers and innovators to increasingly use LLMs as a tool to better understand human behavior, yet the reliability of this tool has not been carefully studied; and $(ii)$ AI algorithms and agents are increasingly being deployed for various tasks in place of humans, yet their performance and reliability remain largely unknown, causing challenges in design efficiency and risk management.

Regarding $(i)$, several studies discuss the potential of GenAI in advancing social science research,

highlighting its ability to enhance research design, experimentation, data analysis, as well as agent-based modeling of complex activities (Charness, Jabarian, and List, 2023; Korinek, 2023; Bail, 2024). These studies take LLMs as a neutral research tool, implicitly treating their responses as unbiased. Our paper challenges this assumption: we systematically study the behavior of LLMs by leveraging the knowledge from the cognitive psychology literature and the experimental economics studies, and we document the behavioral biases that LLMs exhibit——some human-like, others unique to GenAI. Understanding these biases is critical for evaluating GenAI's role in studying human behavior, as they may affect the reliability of LLM-based experiments and simulations. Understanding the behavior of AI agents is also helpful for addressing (*ii*), as the insights can guide the ways in which societies utilize AI technologies while controlling their risks.

Along this line, a new strand of research examines LLMs' performance for tasks previously assigned to humans. Chen et al. (2023) find that, in multiple domains of individual decision making, GPT-3.5 Turbo exhibits a higher degree of economic rationality and a lower degree of choice heterogeneity compared to human participants. Mei et al. (2024) show that GPT-4 exhibits behavioral traits in games that are similar to those from typical human participants. Chen et al. (2025) document that, when forecasting future returns of individual stocks, LLMs manifest biased beliefs that are commonly observed among human participants. Bowen et al. (2025) document that, in mortgage underwriting, loan approvals and denials recommended by LLMs exhibit strong racial biases, which can be mitigated by prompts that require unbiased decisions. Finally, Ouyang, Yun, and Zheng (2024) study how risk preferences of LLMs in financial settings can be modulated by techniques that align LLM behavior with human ethical standards. Together, these studies suggest that LLM behavior is sometimes similar to human behavior, but not always, and is sensitive to prompt framing, training data, and model architecture.

Compared to the studies mentioned above and the broader literature that analyzes LLM performance or algorithmic biases, our work is among the first to advocate studying the behavioral economics of AI as a new research field and treating GenAI agents as a new species. Our work is also more systematic and comprehensive in several important ways. Instead of focusing on specific cases—that is, either a single LLM or isolated aspects of LLM behavior—we systematically document behavioral biases across multiple prominent LLM families; and within each family, we explore

both cross-sectional and time-series variations in LLM responses.[5] When documenting biases, we draw on both the cognitive psychology literature and the experimental economics studies, and we cover both experimental questions that study the psychology of preferences and those that study the psychology of beliefs.[6] Our exploration of debiasing methods is also novel: we compare different methods and propose new ones, with relevance for interventions that aim at reducing biases in real-world settings. Overall, our work lays the foundation for a comprehensive documentation of LLMs' behavioral biases and a systematic exploration of debiasing methods, adding to the nascent literature that calls for LLM evaluations.[7]

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

## 2. Experimental Design

This section describes the experimental design. First, we discuss the selection of questions that study either the psychology of preferences or the psychology of beliefs. Next, we discuss the selection of LLMs. Finally, we discuss our design of API prompts that allow us to systematically collect answers to the experimental questions from the LLMs.

### *2.1. Selection of Experimental Questions*

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

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

[6] Our approach is consistent with the one advocated in Binz and Schulz (2023) and Shiffrin and Mitchell (2023): treating a LLM as a subject in a psychology experiment and studying its responses can be helpful for understanding the LLM's mechanisms of reasoning and decision making.

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

While the traditional theories serve as a rational benchmark for economic studies, decades of research from cognitive psychology casts 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):

> "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 0.5 probability and winning zero with 0.5 probability, and (500) means winning \$500 with certainty. For this question, the majority of participants would choose option B. Then, the same set of participants are 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 0.5 probability and losing zero with 0.5 probability, and (–500) means losing \$500 with certainty. For this question, the majority of participants would choose option C.

It is easy to check 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 such, the same participant choosing option B from the first question and then 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 preferences 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 a prompt design that we describe in Section 2.3; in other words, we replace a human participant by a LLM. This approach allows us to systematically document the behavioral biases of LLMs and compare LLM behavior with human behavior. Table 1 provides a summary of all the experimental questions that our paper currently studies.

[Place Table 1 about here]

Two observations are worth noting. First, for each question in Table 1, a 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, and a non-human-like response that is neither rational nor human-like. Second, Table 1 covers the 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 when making sense of investor behavior, firm behavior, and asset prices observed in financial markets.

Compared to the cognitive psychology literature, a more recent literature from experimental economics studies human behavior by designing and conducting experimental tasks that are more closely tied to real-world economic and financial settings. To further broaden the scope of our analysis, we also collect LLM responses to a set of recent experimental tasks from this literature. In particular, we follow Afrouzi et al. (2023) by asking the LLMs to first observe a sequence of past realizations of a random variable $x_t$ and then forecast its future realizations; 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 $\rho$ measures the persistence of the process and $\epsilon_t$ is an i.i.d. Gaussian random variable.

As in Afrouzi et al. (2023), we consider three experiments. In the baseline experiment, a LLM is endowed with the knowledge that the evolution of $x_t$ is 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 then asked, at time 41, to forecast the next two outcomes, $x_{42}$ and $x_{43}$; such a procedure continues until the LLM observes 44 past realizations of $x_t$ and is then asked, at time 44, to forecast the next two outcomes, $x_{45}$ and $x_{46}$. The second and third experiments each serve as a variant to the baseline experiment. The second experiment is identical to the baseline experiment, except that, at each time $t$, the LLM is asked to forecast $x_{t+1}$ and $x_{t+5}$; for example, at time 40, the LLM first

observes 40 past realizations of $x_t$, ranging from $x_1$ to $x_{40}$, and is then asked to forecast $x_{41}$ and $x_{45}$. The third experiment is identical to the baseline experiment, except that the LLM is now endowed with more detailed knowledge that the evolution of $x_t$ is "a fixed and stationary AR(1) process: $x_t = \mu + \rho x_{t-1} + \epsilon_t$, with a given $\mu$, a given $\rho$ in the range [0,1], and an $\epsilon_t$ that is an i.i.d. random shock."

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

### *2.2. Selection of LLMs*

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

[Place Table 2 about here]

These twelve models differ both across and within families, especially along the following three dimensions: size of the training data, design of the model architecture, and the reinforcement learning algorithm. In terms of the training data, newer models are trained on more data compared to older models; for example, Meta Llama explains that the training dataset of Llama 3 consists of over 15 trillion tokens, while Llama 2 consists of 1.8 trillion tokens only.[8]

[8]For more details about the characteristics of Meta Llama 3, see: https://ai.meta.com/blog/meta-llama-3/. For the other three families, the exact training data are not often disclosed publicly; nonetheless, newer models in general tend to be trained on more data. Brown et al. (2020) explain that OpenAI used around 500 billion tokens to train

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 in the model architecture vary significantly from one model to another. For example, among the older generation of models, the largest is Claude 2, which has an estimate of 200 billion parameters and a context window of approximately 100,000 words (tokens), whereas the smallest is Llama 2 with 70 billion parameters and a context window of approximately 4,000 words. Second, within each family, the model architecture has evolved significantly between 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 can 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 the reinforcement learning algorithm, each model relies on a different implementation of the Reinforcement Learning from Human Feedback (RLHF) algorithm to align its answers with human preferences. For example, Anthropic Claude combines RLHF with a method called Constitutional AI, which aligns the model behavior with human principles of helpfulness, harmlessness, and honesty (Bai et al., 2022).

### *2.3. Prompt Design*

We collect LLM responses to each of our experimental questions through an application programming interface (API). The API takes as input a "prompt," which is a text file submitted to a LLM in order to receive a response back. Below, we describe the prompt design that allows for

GPT 3.5. The exact number of tokens used to train GPT-4 is not known, although unofficial sources claim that OpenAI used around 13 trillion tokens to train GPT-4; for more information on GPT-4's training data, please visit: https://semianalysis.com/2023/07/10/gpt-4-architecture-infrastructure/.

[9] Mixture-of-experts architectures use a "router," otherwise called a 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, while GPT-3.5 Turbo uses a single-expert architecture with 175 billion parameters, 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] Two commonly used compression techniques are: quantization, which reduces parameter precision, and pruning, which removes the less important connections from the neural network.

[11] In the case of Meta Llama, the architecture is very similar between Llama 3 8B and Llama 3 70B: the two models are trained on the same dataset and they use similar architectures. However, Llama 3 8B uses fewer parameters compared to Llama 3 70B (8 billion versus 70 billion).

elicitation of desired responses from LLMs.

A proper prompt needs to satisfy two requirements. First, it must instruct the LLMs to provide standardized responses for subsequent analysis. Second, it must contain questions that are similarly phrased compared to the original experimental questions used to study human behavior. Given these two requirements, Fig. 1 provides an example—the prompt we use to elicit LLM responses to a question that Kahneman and Tversky (1979) design for documenting diminishing sensitivity as a key element of prospect theory.

[Place Fig. 1 about here]

Fig. 1 shows that a prompt is structured in three parts; this applies to all experimental questions listed in Table 1. The first part contains a general instruction that asks a LLM to consider a specific set of experimental scenarios; in Fig. 1, this part starts from "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 starts from "The output should be" and ends with "} ‘‘‘".[12] The third part is the main element of the prompt. It contains the precise experimental questions designed by psychologists to study human behavior; in Fig. 1, this part starts from "Scenario A:" and ends with "calculations)." from Scenario B. At the end of describing each scenario, further instructions are given to the LLM, to make sure that it provides the set of responses that we elicit.

Two observations are worth noting. First, with our prompt design, a LLM response typically contains, for each experimental scenario, four different parts: choice, confidence, explanation, and reasoning. Here, "choice" refers to an explicit choice made by the LLM—for example, whether the LLM accepts or turns down a risky gamble.[13] "Confidence" refers to the confidence level the LLM assigns to its choice using a score between 0 and 1. "Explanation" refers to a brief explanation that the LLM provides to justify its choice. And "reasoning" asks for choosing between two reasoning types: type "A" corresponds to reasoning that is based more on intuitive thinking, while type "B"

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

[13]Instead of eliciting a "choice" between multiple options, Question 10 (regarding "base rate neglect") and Question 12 (regarding "gambler's fallacy") ask for an estimate of a probability; Question 14 (regarding "anchoring") asks for an estimate of a percentage number.

corresponds to reasoning that is based more on analytical thinking and calculations.[14] Second, many experimental questions we examine document behavioral biases by having the *same* participant provide responses in different scenarios; for example, as discussed in Section 2.1, Kahneman and Tversky (1979) document the diminishing sensitivity element of prospect theory by having the same human participant answer two different questions—one that frames lottery payoffs as gains and the other that frames lottery payoffs as losses. Such experimental questions require a "within-subject" design that allows us to think of a LLM as a participant and elicits its responses in 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 a sentence that instructs the LLM to "treat each scenario as completely separate from the other."[15] The Internet Appendix contains prompt design for all sixteen questions described 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 a separate prompt that implements the Afrouzi et al. (2023) experiments described in Section 2.1. First, these experiments require not only textual inputs but also *graphical* inputs: participants are presented with both textual instructions and figures that plot past realizations of a random variable. To satisfy this requirement, a LLM needs to support graphical inputs; this leads to the exclusion of six LLM platforms.[16] For the remaining LLMs that support graphical inputs, we follow platform-specific guidelines when uploading figures.[17] Second, the LLMs do not always provide precise forecasts that we elicit when they are presented with figures; sometimes, they refuse to respond. To address this issue, 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 do so." Third and finally, the Afrouzi et al. (2023) experiments require that the same individual makes multiple rounds of forecasts; each round depends on textual and graphical inputs presented up to that point in time. To enforce such sequential dependence, we implement a sequence of API calls. In particular, for each call, we

[14] Our current analysis uses the "choice" part only. We plan to use the other three parts in future iterations.

[15] LLM responses can be random—asking the same LLM an identical question multiple times can yield varying responses. As such, we find it plausible to view each API call as an individual participant.

[16] 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.

[17] Google Gemini directly processes figures that are 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.

feed the entire conversation history—including all previous prompts and responses—into the LLM, hence 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; we treat these models as our benchmark models; and we analyze how they respond to the questions from psychology, 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 the LLM responses to preference-based questions and their responses to belief-based questions. We then explore the heterogeneity in LLM responses across LLM families, model generations, and model scales. Finally, we examine the LLM responses to questions from the experimental economics tasks that 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 from 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; in other words, for each LLM, we iterate over each question 100 times. Each iteration consists of an API call submitted to the model, whereby the prompt for the specific question is provided as an 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.[18] Note that setting the temperature parameter to zero results in deterministic outputs, while higher values increase the randomness in LLM responses.[19]

[18]The range for the temperature parameter varies across platforms: for ChatGPT and Anthropic Claude, the range is [0, 1]; for Meta Llama, the range is [0, 5]; finally, the range for Gemini 1.0 Pro is [0, 1] and the range for Gemini 1.5 Pro and Gemini 1.5 Flash is [0, 2].

[19]Specifically, for the iterative process of generating each word (token) in a response, the LLM first forms a

We collect and analyze each LLM response, categorizing it into one of the three groups: rational, human-like, or other. A response is categorized as rational if a LLM's choice or estimate aligns with that of an agent who has rational preferences and rational beliefs; a response 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 a response falls into the category of "other" 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 a LLM makes the same choices, we categorize such a response 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, we categorize such a response as "other."

[Place Fig. 2 and Table 3 about here]

Fig. 2 summarizes the responses obtained from the four benchmark models of GPT-4, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3 70B. For each model, we categorize the sixteen experimental questions from the cognitive psychology literature into two groups: preference-based questions (left panel) and belief-based questions (right panel). The results are presented using bar charts that depict the proportion of responses categorized as rational (blue), human-like (red), or other (gray).[20] Table 3 provides 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%.

Two important observations are worth noting. First, the majority of the LLM responses fall into either the rational category or the human-like category, with the responses registered as "other" in just a few cases. Specifically, for GPT-4, "other" responses are observed only in Question 3, which

probability distribution over all possible tokens in its dictionary and then draws the next token from this distribution. The temperature parameter reshapes the distribution: a higher temperature parameter makes the distribution more uniform, hence 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 only; and top-$p$ sampling retains a subset of the top-$k$ most probable tokens whose cumulative probability, when normalized using the total probability of the top-$k$ most probable tokens, exceeds the threshold of $p$. In our analysis, we set $k$ to its default value of 50 and $p$ to its default value of 0.9.

[20] 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.

pertains to the probability weighting element of prospect theory. For Claude 3 Opus, "other" responses are observed 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 are observed in one preference-based question only—Question 3 on probability weighting. 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: the 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 the LLMs produce human-like responses more than 50% of the time. Specifically, Gemini 1.5 Pro has the majority of responses categorized as human-like in five out of six questions; Claude 3 Opus has the majority of responses categorized as human-like in four out of six questions; and GPT-4 and Llama 3 70B have the majority of responses categorized as human-like in three out of six questions. For most belief-based questions, the LLMs produce rational responses more than 50% of the time. Specifically, Gemini 1.5 Pro has the majority of responses categorized as rational in ten out of ten questions; both GPT-4 and Claude 3 Opus have the majority of responses categorized as rational in eight out of ten questions; and Llama 3 70B has the majority of 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 our analysis to examine all twelve models. These include three models for each LLM family: (*i*) the benchmark model that is highly advanced and large-scale, (*ii*) a highly advanced model with a smaller scale, and (*iii*) a large-scale model of an older generation. We begin by examining variations in responses across the four LLM families. Then, controlling for LLM family fixed effects, we analyze how variations in model generation and model scale influence the 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 variations among the four benchmark models. For the 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 examine the heterogeneity in responses across the four 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. For 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. For 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 three LLM families of Claude, Gemini, and Llama, with the LLM family of GPT serving as the omitted baseline.

[Place Table 4 about here]

Table 4 reports the marginal effects from the above probit regressions, where each reported coefficient represents the change in the predicted probability of observing an outcome $Y_{iqk}$ of one that is associated with changing the LLM from GPT to each of Claude, Gemini, and Llama. Consistent with the heterogeneity observed from Fig. 2 across the LLM families, for the preference-based questions, Gemini models are 22.9% less likely to produce a rational response, compared to 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, compared to GPT models; this effect is significant at the 5% level. Moreover, the responses from Claude or Llama models to the preference-based

questions are statistically similar to those from GPT models.

Also consistent with the heterogeneity observed from Fig. 2 across the LLM families, for the belief-based questions, Llama models are 25.0% less likely to produce a rational response, compared to GPT models; this effect is significant at the 5% level. Llama models are 21.0% more likely to produce a human-like response, compared to GPT models; and this effect is also significant at the 5% level. Finally, the responses from Claude or Gemini models to the belief-based questions are statistically similar to those from GPT models. Taken together, the findings from Table 4 highlight meaningful LLM family-level differences in responses to experimental questions drawn from cognitive psychology. In our subsequent analyses on the heterogeneity across model generations and model 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. The evolutions of model generation and model scale capture key aspects of LLM development, including improvements of 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 on LLM responses, we compare large-scale models with smaller-scale ones of the same generation. For both comparisons, we control for LLM family fixed effects.

[Place Fig. 3 about here]

We begin by presenting graphical evidence on the differences in LLM responses across model generations and model 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 variations. 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 in Fig. 3 reveal a striking contrast between the LLM responses to the preference-based questions and their responses to the belief-based questions. For the preference-based ques-

tions, the left panel of Fig. 3 shows that, as the LLMs become more advanced or larger, the number of questions that receive predominantly rational responses tends to decrease, while the number of questions receiving predominantly human-like responses increases. For the belief-based questions, the right panel of Fig. 3 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 the heterogeneity in LLM responses across model generations and model scales, we estimate a series of probit regressions. In particular, we conduct two analyses. First, to examine the effect of model generation on LLM responses, we restrict our sample to the LLM 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 $k$. For 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. For 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 four advanced models; moreover, $\gamma_f$ captures LLM family fixed effects.

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

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

For 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. For 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 four large-scale models.

[Place Table 5 about here]

Table 5 reports the marginal effects from the above probit regressions, where the reported coefficients represent the change in the predicted probability of observing an outcome $Y_{iqk}$ of one that is associated with either moving from an older model to an advanced model or from a smaller-scale model to a large model. The regression results are by and large consistent with the variations in LLM responses observed from Fig. 3 across model generations and model scales. For preference-based questions, Columns (1) to (4) in Panel A show that, as the 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 the 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. Most of the coefficients reported in Columns (1) to (4) are statistically significant; the only exception is that, as the models become larger, the increase in human-like responses to the preference-based questions is insignificant.

For belief-based questions, Columns (5) to (8) in Panel A show that the 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 the models become larger in scale. All the coefficients reported in Columns (5) to (8) are statistically significant.

In summary, both Fig. 3 and Table 5 show systematic heterogeneity in LLM responses across model generations and model scales. As the 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 the 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 $\mu$, the constant term, to 0 and setting $\sigma$, the standard deviation of $\epsilon_t$, to 20; these parameter values are taken from Afrouzi et al. (2023). For $\rho$, the persistence parameter, we take six values of 0, 0.2, 0.4, 0.6, 0.8, and 1; and for each value, we generate 100 simulated paths. As such, each experiment has a total of 600 different paths.

For a given experiment and a given simulated path, we ask the 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 the LLM has completed five rounds of forecasts.

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 $\rho$, 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 our estimates based on the LLM responses to questions from Experiment 1 of Afrouzi et al. (2023); here, we focus on LLMs' short-term forecasts, with a forecasting horizon $s$ of one. The top panel displays the $\hat{\rho}$ values estimated for each of three baseline models: GPT-4, Claude 3 Opus, and Gemini 1.5 Pro. The bottom panel displays the $\hat{\rho}$ values estimated for each

of the three smaller-scale models: GPT-4o, Claude 3 Haiku, and Gemini 1.5 Flash.[21] For each $\hat{\rho}$ estimate, we also plot its 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]

Fig. 4 gives rise to two observations about the LLM forecasts from Experiment 1. First, the top panel shows that, for the advanced large-scale models of GPT-4, Claude 3 Opus, and Gemini 1.5 Pro, the LLM forecasts are by and large rational; for each model and each $\rho$, the estimated $\hat{\rho}$ is reasonably close to $\rho$. Second, the bottom panel shows that, for the smaller-scale models of GPT-4o, Claude 3 Haiku, and Gemini 1.5 Flash, the LLM forecasts are human-like: consistent with the findings of Afrouzi et al. (2023) on human participants, the persistence $\hat{\rho}$ implied by the LLM forecasts is significantly higher than the true persistence $\rho$; moreover, the difference between $\hat{\rho}$ and $\rho$ is larger for lower values of $\rho$. The comparison between the top panel and the bottom panel reinforces a pattern documented in Section 3.1 using the experimental questions from the psychology literature, namely that larger-scale models tend to generate more rational responses to belief-based questions.

We also examine the LLM forecasts from Experiments 2 and 3: using these forecasts, Fig. 5 plots the perceived persistence $\hat{\rho}$ against the true persistence $\rho$. Specifically, the top panel examines the LLMs' longer-term forecasts $F_{it}x_{t+5}$ from Experiment 2; these are model $i$'s time-$t$ forecasts of $x_{t+5}$. The bottom panel examines the LLMs' short-term forecasts $F_{it}x_{t+1}$ from Experiment 3; in this experiment, the LLMs are provided with more detailed knowledge about the evolution of $x_t$.

[Place Fig. 5 about here]

The comparison between Fig. 4 and Fig. 5 yields two interesting observations. First, for the three advanced large-scale models of GPT-4, Claude 3 Opus, and Gemini 1.5 Pro, the LLMs' longer-term forecasts give rise to human-like responses that are absent from the short-term forecasts: the top panel of Fig. 5 shows that, the persistence $\hat{\rho}$ implied by the LLMs' longer-term forecasts is significantly higher than the true persistence $\rho$; moreover, the difference between $\hat{\rho}$ and $\rho$ is larger for lower values of $\rho$. In comparison, as discussed before, the top panel of Fig. 4 shows that the

[21] Llama models do not support graphical inputs, so they are excluded from this analysis.

LLMs' short-term forecasts are by and large 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 provision of detailed information about the data generating process can be counterproductive: it gives rise to more human-like biases in LLM responses. Interestingly, this novel result is specific to LLMs; Afrouzi et al. (2023) find that human participants' forecasts are unaffected by the provision of more information about the evolution of $x_t$. Section 4 provides more discussion of this finding.

## 4. Correcting LLM Biases

Section 3 has documented that, for many preference-based questions and belief-based questions, the LLM responses exhibit behavioral biases. In this section, we explore role-priming methods—instructing a LLM to view itself as a certain type of individual—as well as debiasing techniques that aim at correcting the observed LLM biases.

We begin by discussing how role priming affects the LLM responses. Here, we consider two versions: the first version instructs the LLMs to view themselves as rational investors; the second version instructs the LLMs to view themselves as real-world retail investors. We implement each version by adding one sentence at the beginning of the prompt. 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 first version and "When answering questions below, please think of yourself as a real-world retail investor who makes economic and financial decisions." for the second version.

[Place Table 6 about here]

Table 6 presents the effects of role priming on LLM behavior. Panel A reports the treatment effects of priming the LLMs to be a rational investor. Averaged across the twelve LLMs, such role priming increases rational responses by 4.3% for the preference-based questions (significant at the 5% level) and increases rational responses by 3.3% for the belief-based questions (significant at the 10% level).[22] Panel B reports the treatment effects of priming the LLMs to be a real-world retail

[22]Here we report the treatment effects with model fixed effects—acknowledging the differences across the twelve LLMs—included as a control. Without this control, the treatment effects are by and large similar.

investor. Averaged across the twelve LLMs, such role priming reduces rational responses by 3.9% for the preference-based questions (significant at the 5% level); and it does not cause a significant change in the LLM responses to the belief-based questions. Table 6 suggests that instructing the LLMs to behave as rational investors is effective in reducing biases; in other words, such role priming can be used as a debiasing technique.

Table 7 explores two more debiasing techniques. The first technique combines the sentence that primes the LLMs to be rational investors with the provision of a detailed four-step procedure that guides the LLMs to rationally choose a course of action 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 the LLMs to be rational investors with the provision of a summary of the key findings from Kahneman and Tversky (1979) that describe biased human behavior. The summary is generated by first uploading the .pdf form of the original Kahneman and Tversky (1979) paper to an interactive GPT-4o chat box and then asking for 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 have the LLMs avoid making the same mistakes. Therefore, 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 7 about here]

Table 7 compares the baseline debiasing technique of simply priming the 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 proba-

bility weighting.[23] Table 7 shows that the provision of the four-step procedure that guides the 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 26% and increases human-like responses by about 18%. Taken together, these findings suggest that provision of more information—even if genuinely useful for decision making—is not always useful in correcting LLM biases: information overload may hinder a LLM's ability to provide rational responses. This is consistent with our finding from Section 3.3 that provision of more information about the data generating process in the Afrouzi et al. (2023) experiments gives rise to more human-like biases in LLM responses.

## 5. Conclusion

Artificial intelligence, especially generative AI epitomized by LLMs, has become increasingly important in social and economic activities. Our paper calls for systematically studying the behavioral economics of AI, and as a starting point, we examine the behavior of four prominent families of LLMs—ChatGPT, Anthropic Claude, Google Gemini, and Meta Llama—by leveraging the experimental designs used in the cognitive psychology literature and the experimental economics studies.

We document systematic patterns in the behavioral biases that LLMs exhibit. For experimental questions that study human preferences, the LLMs' responses become increasingly irrational and human-like as we move towards more advanced models or models with a larger scale. For questions that study human beliefs, however, the most advanced large-scale LLMs generate responses that are by and large rational; here, we examine belief-based questions from both the psychology literature and the experimental economics studies. Moreover, we observe significant heterogeneity in responses across the four families of LLMs.

We also explore role-priming and debiasing methods that might affect LLM behavior. In particular, a prompt that instructs LLMs to behave as rational investors who make decisions according to the Expected Utility framework is effective in reducing biases; a prompt that instructs LLMs

[23] We focus on the three prospect theory-related questions for two reasons. First, the LLM responses to these questions are often irrational, suggesting that there is room for debiasing. Second, one debiasing technique described above involves the provision of prospect theory's key findings. Such information will mostly likely affect the LLMs' responses to prospect theory-related questions.

to behave as real-world retail investors leads to less rational responses when the LLMs answer preference-based questions. Finally, provision of bias-reducing information—either a detailed procedure that guides the LLMs to rationally choose a course of action under the Expected Utility framework or a summary of key findings from Kahneman and Tversky (1979) that describe biased human behavior—is not found useful in reducing LLM biases.

## References


Afrouzi, H., Kwon, S. Y., Landier, A., Ma, Y., Thesmar, D., 2023. Overreaction in Expectations: Evidence and Theory. The 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.

Bose, D., Cordes, H., 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.

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.

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.

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.

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.

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: *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 the LLMs' responses to a question that Kahneman and Tversky (1979) design for documenting diminishing sensitivity as a key element of prospect theory.

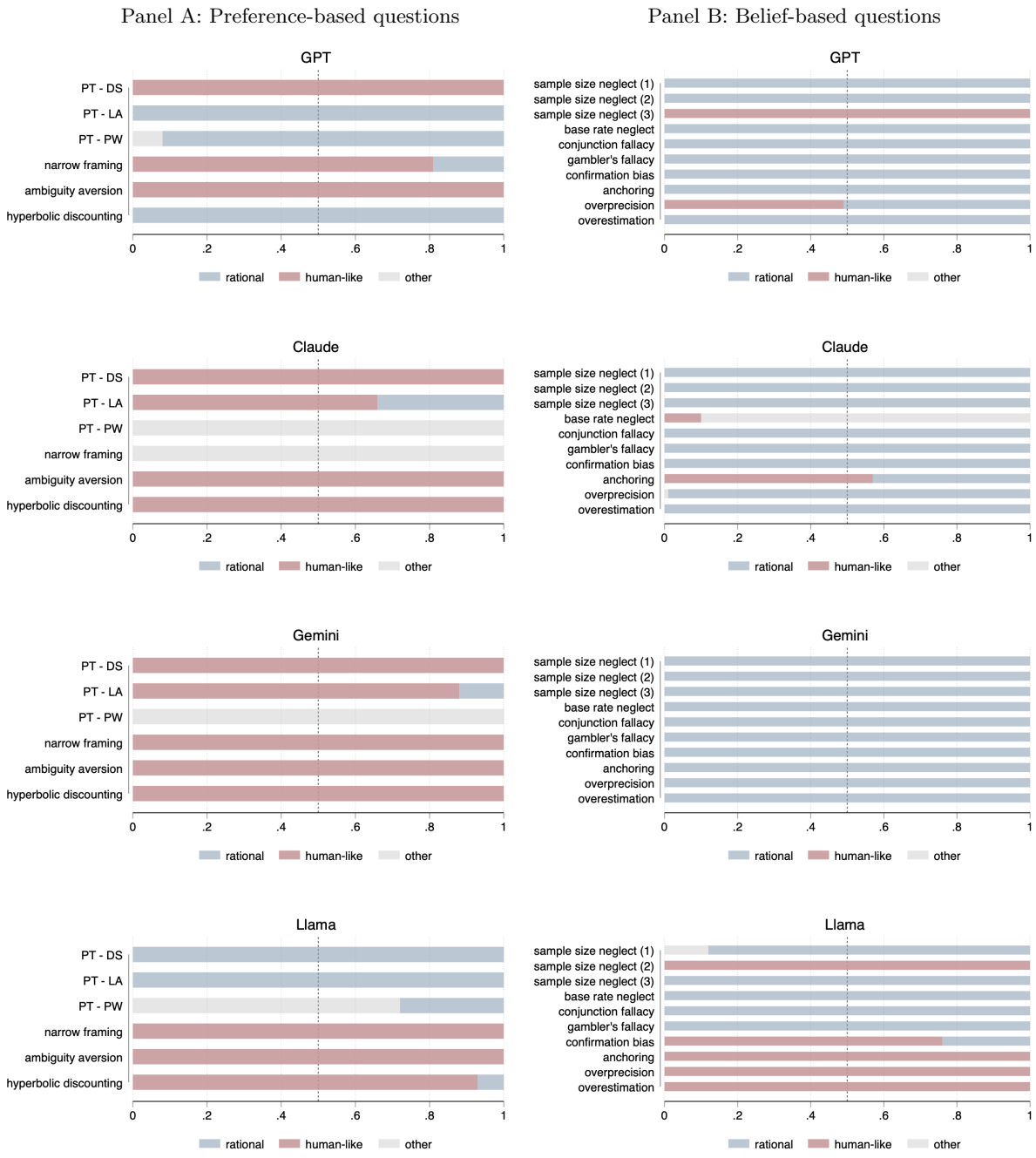


**Fig. 2. Proportion of LLM responses: Advanced large-scale models.**

This figure plots the proportion 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. The right panel presents results for the ten belief-based questions.

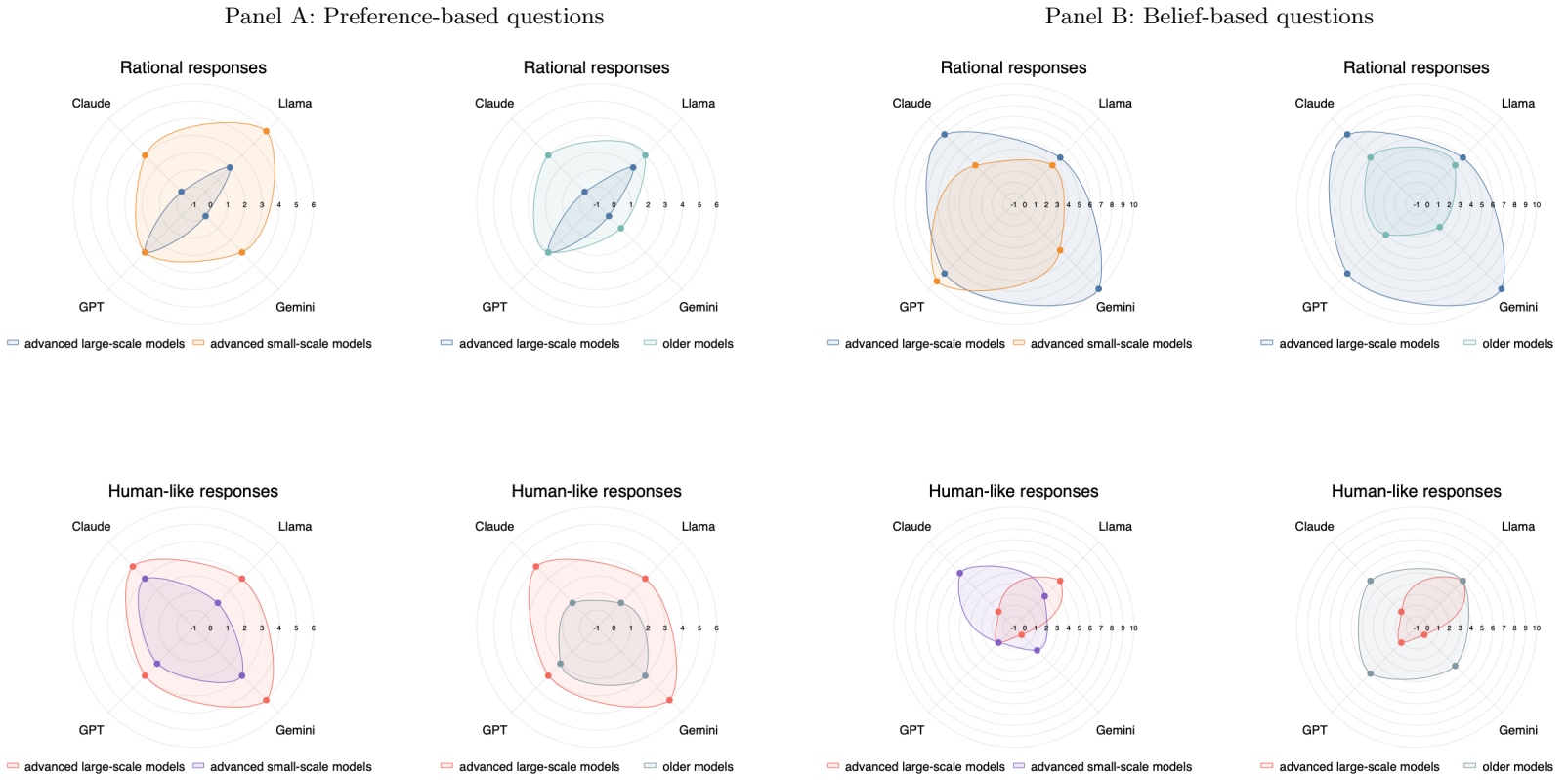




**Fig. 3. Heterogeneity in LLM responses across model generations and model scales.**

This figure presents radar charts that compare the number of questions that receive predominantly rational responses (top row) or human-like responses (bottom row) across different LLMs, 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.

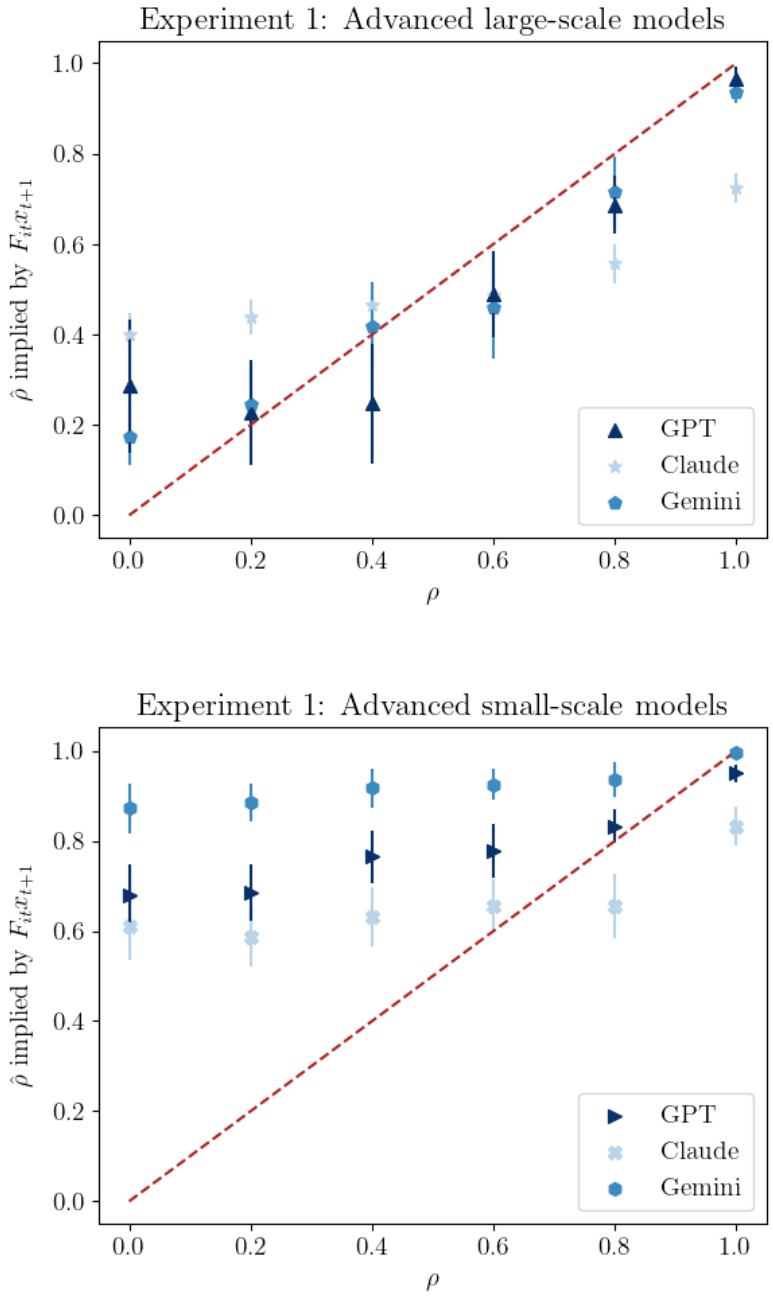


**Fig. 4. LLM forecasts: Experiment 1 of Afrouzi et al. (2023).**

The figure plots the perceived persistence $\hat{\rho}$ against the true $\rho$. Here, $\hat{\rho}$ is estimated using the LLMs' forecasts from Experiment 1 of 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; 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 its 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).

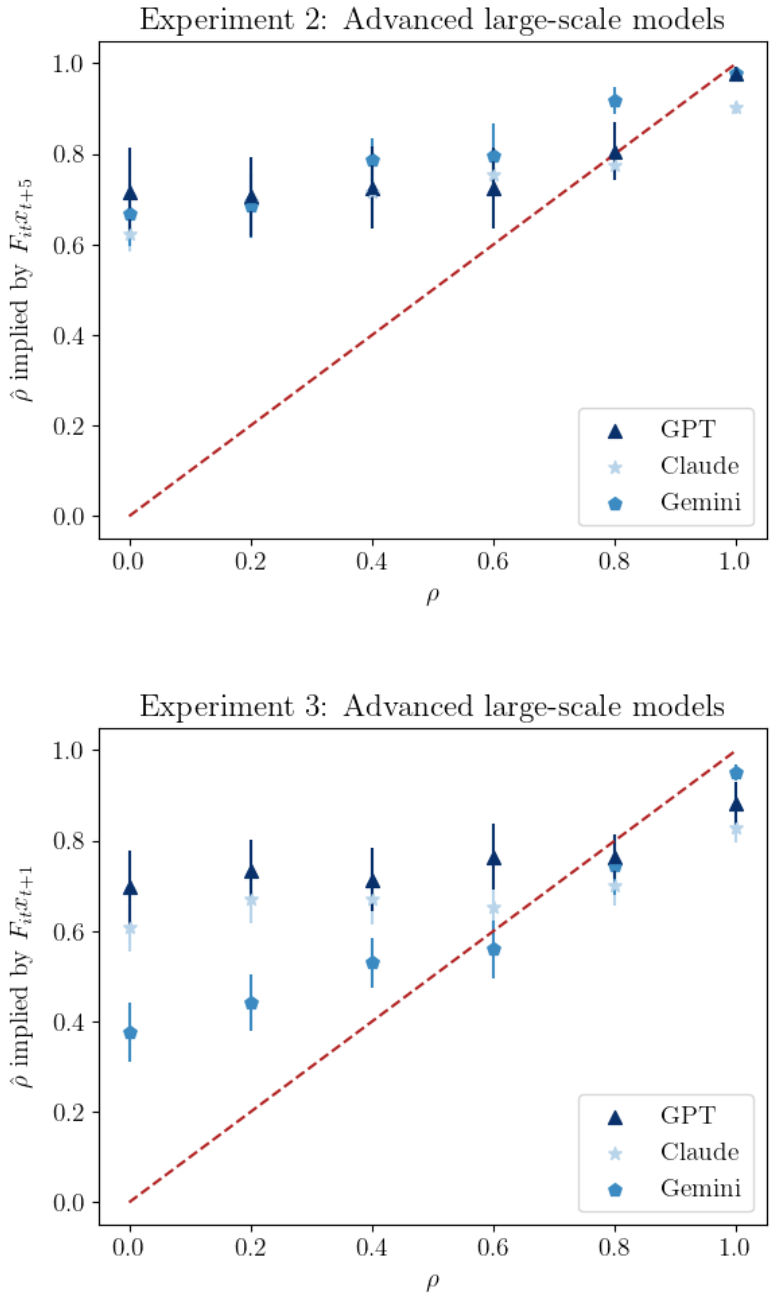


**Fig. 5. LLM forecasts: Experiments 2 and 3 of Afrouzi et al. (2023).**

The figure plots the perceived persistence $\hat{\rho}$ against the true $\rho$. Here, $\hat{\rho}$ is estimated using the LLMs' forecasts from Experiments 2 and 3 of Afrouzi et al. (2023): the top panel examines the LLMs' forecasts from Experiment 2; the bottom panel examines the LLMs' forecasts from Experiment 3. For both panels, we 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 its 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 provides a summary of the sixteen experimental questions we examine. These questions are drawn from the cognitive psychology literature. Specifically, Question 1 is based on Problems 11 and 12 of 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′ of Kahneman and Tversky (1979) (page 281). Question 4 is a modified version of Problem 10 from 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 of Well, Pollatsek, and Boyce (1990) (page 297). Question 9 is based on Problem 10 of 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). Question 16 follows the procedure discussed on pages 508 to 509 of Moore and Healy (2008) to document overestimation.

Panel A: A list of 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: A list of 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 provides a description of the twelve LLMs that we examine. We group these models by four LLM families: ChatGPT, Anthropic Claude, Google Gemini, and Meta Llama. For each family, we consider the advanced and large-scale models as our 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—and their predecessors—GPT-3.5 Turbo, Claude 2, Gemini 1.0 Pro, and Llama 2 70B. RLHF and RLAI are the abbreviations for "Reinforcement Learning from Human Feedback" and "Reinforcement Learning from AI," respectively. MMLU is the abbreviation for "Massive Multitask Language Understanding" and it provides a benchmark score for evaluating the capabilities of LLMs. Vision indicates whether a model supports graphical inputs or not.

| Model | Release year | Size (number of parameters) | Data (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. Panel B presents results for the ten belief-based questions. The abbreviations of "PT - DS," "PT - LA," and "PT - PW" are for "prospect theory - diminishing sensitivity," "prospect theory - loss aversion," and "prospect theory - probability weighting," respectively. The 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 the marginal effects from the probit regressions specified by:

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

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 three LLM families of Claude, Gemini, and Llama, with the LLM family of GPT serving as the omitted baseline category. The reported coefficients represent the change in the predicted probability of observing an outcome $Y_{iqk}$ of one that is associated with changing the LLM from GPT to each of Claude, Gemini, and 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 the marginal effects from the probit regressions specified in equations (3) and (4) of the main text. For Columns (1), (2), (5), and (6), the dependent variable is a binary variable that takes the value of one if a 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 a LLM response is classified as human-like, and zero other wise. Regressions in Columns (1) to (4) are for preference-based questions and regressions in Columns (5) to (8) are for belief-based questions. Panel A compares advanced large-scale models with older models. In this case, we restrict the sample to the LLM responses from either the advanced large-scale models or the older models; the key independent variable is an indicator for the advanced models, with the older models serving as the baseline. Panel B compares large-scale models with smaller ones. In this case, we restrict the sample to LLM responses from either the advanced large-scale models or the advanced smaller-scale models; 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. Treatment effects of role-priming prompts.**

This table reports the 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 a 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 a 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 the twelve LLMs. Panel A restricts the sample to the LLM responses generated using either the baseline prompt or a treatment prompt that primes the LLMs to be rational investors; Panel B restricts the sample to the LLM responses generated using either the baseline prompt or a treatment prompt that primes the 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 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 7. Comparison of debiasing techniques: Prospect theory-related questions.**

This table reports the 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 a 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 a LLM response is classified as human-like, and zero otherwise. Regressions are estimated using the LLM responses to prospect theory-related questions only; each regression uses responses from all the twelve LLMs. Panel A restricts the sample to the LLM responses generated using either the baseline prompt or a treatment prompt that primes the LLMs to be rational investors; Panel B restricts the sample to the LLM responses generated using either the baseline prompt or an instruction-based prompt that combines the sentence that primes the LLMs to be rational investors with the provision of a detailed four-step procedure that guides the LLMs to rationally choose a course of action; Panel C restricts the sample to the LLM responses generated using either the baseline prompt or a knowledge-enrichment prompt that combines the sentence that primes the LLMs to be rational investors with the provision of a summary of the key findings from Kahneman and Tversky (1979) that describes biased human behavior. For all three panels, the key independent variable is an indicator for the treatment prompt, with the baseline prompt serving as the omitted 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) |
| 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 | –0.0617 | –0.0596 | 0.0614 | 0.0605 |
| | (0.079) | (0.077) | (0.081) | (0.084) |
| Model FE | No | Yes | No | Yes |
| Observations | 7,200 | 7,200 | 7,200 | 6,600 |
| Pseudo *R*-squared | 0.003 | 0.204 | 0.004 | 0.184 |

| 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 | –0.269*** | –0.263*** | 0.185* | 0.185* |
| | (0.069) | (0.065) | (0.111) | (0.106) |
| Model FE | No | Yes | No | Yes |
| Observations | 7,196 | 7,196 | 7,196 | 7,196 |
| Pseudo *R*-squared | 0.054 | 0.222 | 0.029 | 0.136 |

# Internet Appendix for
# "Behavioral Economics of AI: LLM Biases and Corrections"

PIETRO BINI, LIN WILLIAM CONG, XING HUANG, and LAWRENCE J. JIN

This Internet Appendix contains the following two sections:

- Section I: Additional Figures
- Section II: Prompt Design

# I. Additional Figures

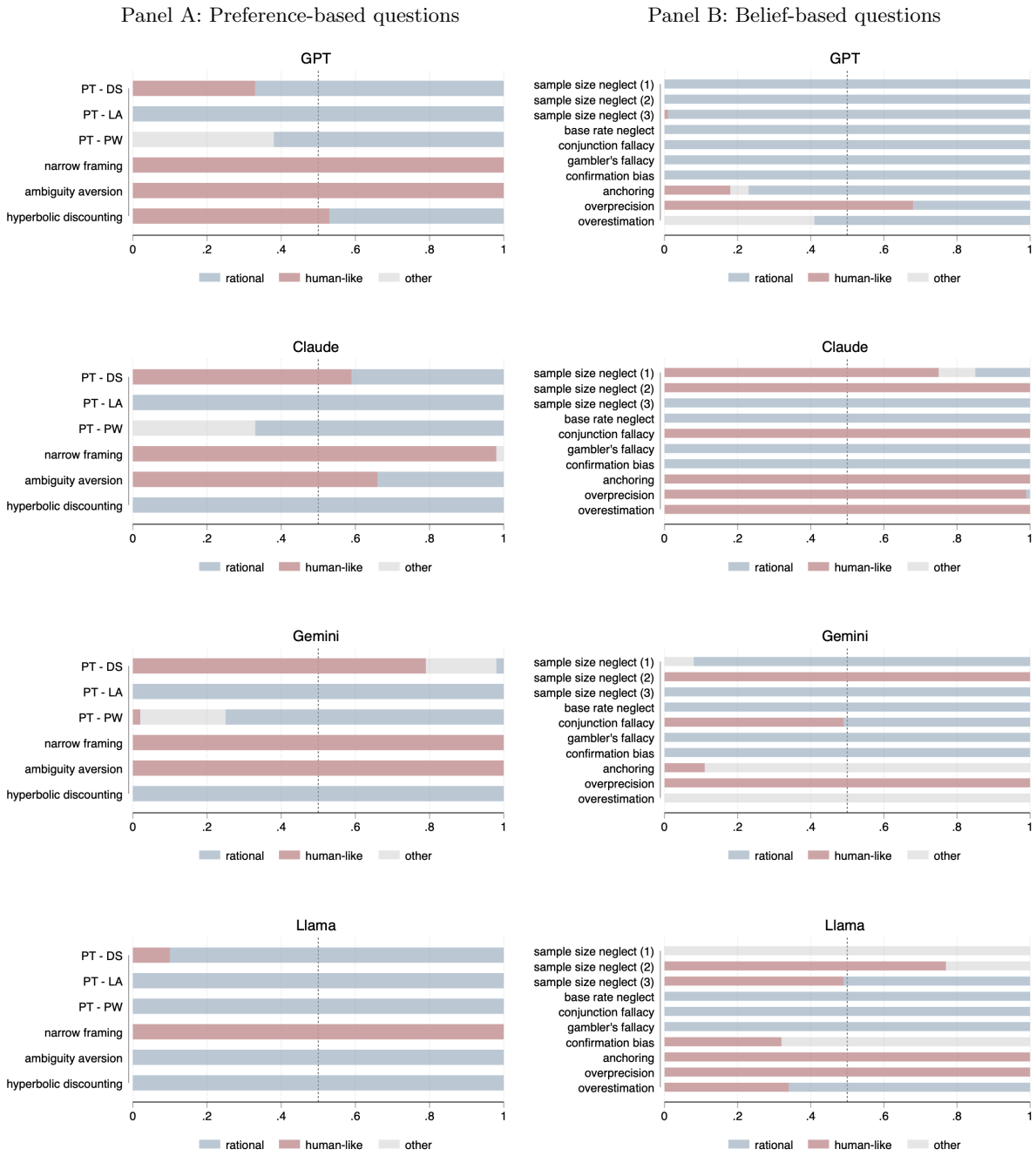


**Fig. IA.1. Proportion of LLM responses: Advanced small-scale models.**

This figure plots the proportion of LLM responses categorized as rational (blue), human-like (red), or other (gray), for the four advanced small-scale LLMs: GPT-4o, Claude 3 Haiku, Gemini 1.5 Flash, and Llama 3 8B. The left panel presents results for the six preference-based questions. The right panel presents results for the ten belief-based questions.

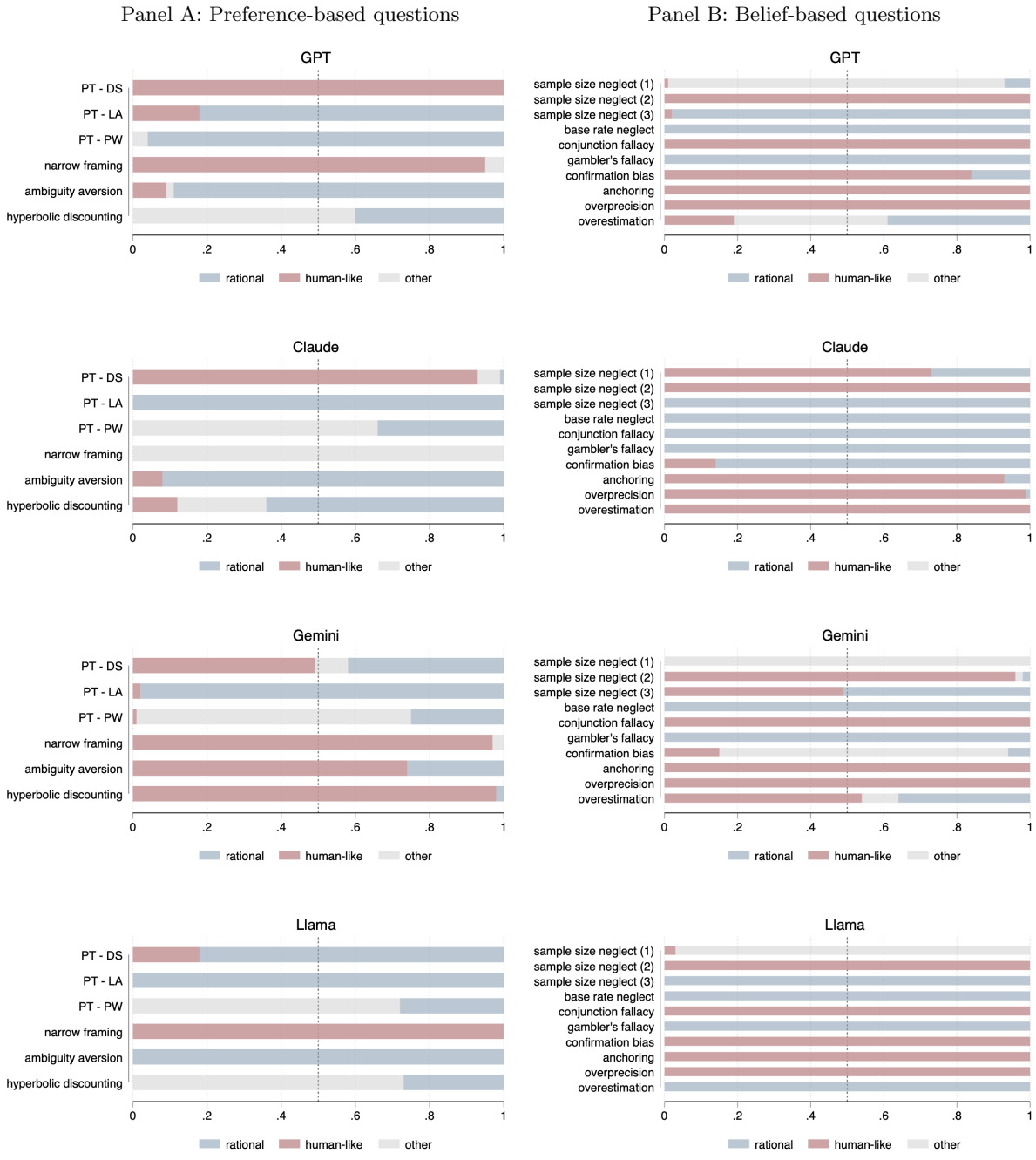


**Fig. IA.2. Proportion of LLM responses: Older models.**

This figure plots the proportion of LLM responses categorized as rational (blue), human-like (red), or other (gray), for the four older versions of LLMs: GPT-3.5 Turbo, Claude 2, Gemini 1.0 Pro, and Llama 2 70B. The left panel presents results for the six preference-based questions. The right panel presents results for the ten belief-based questions.

## II. Prompt Design

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. IB.1. Prompt for question 1: Diminishing sensitivity of prospect theory.**

This figure presents a prompt that elicits the LLMs' responses to a question that Kahneman and Tversky (1979) design for documenting diminishing sensitivity as a key element of prospect theory.

Instructions:
Consider the following question and respond according to the template provided. 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
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

Question:
Would you accept or turn down a 50:50 bet to win $110 or lose $100?

Response Format:
Please answer as shown above. Indicate the choice you prefer ("Accept" or "Turn down"), 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. IB.2. Prompt for question 2: Loss aversion of prospect theory.**

This figure presents a prompt that elicits the LLMs' responses to a question discussed in Barberis and Thaler (2003) that documents loss aversion as a key element of prospect theory.

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:
Please consider the following scenario. Choose between the following two options: option A ($5,000, 0.001), meaning receiving $5,000 with 0.001 probability and receiving zero with 0.999 probability, versus option B ($5), meaning receiving $5 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, choose between the following two options: option A (–$5,000, 0.001), meaning losing $5,000 with 0.001 probability and losing zero with 0.999 probability, versus option B (–$5), meaning losing $5 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. IB.3. Prompt for question 3: Probability weighting of prospect theory.**

This figure presents a prompt that elicits the LLMs' responses to a question that Kahneman and Tversky (1979) design for documenting probability weighting as a key element of prospect theory.

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:
Please consider the following scenario. Imagine that you are about to purchase a jacket for $125 and a calculator for $15. The calculator salesman informs you that the calculator you wish to buy is on sale for $10 at the other branch of the store, located 5 minutes drive away. Would you make the trip to the other store? Please answer as shown above. Indicate the choice you prefer ("Yes" making the trip or "No" not making the trip), 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, imagine that you are about to purchase a jacket for $15 and a calculator for $125. The calculator salesman informs you that the calculator you wish to buy is on sale for $120 at the other branch of the store, located 5 minutes drive away. Would you make the trip to the other store? Please answer as shown above. Indicate the choice you prefer ("Yes" making the trip or "No" not making the trip), 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. IB.4. Prompt for question 4: Narrow framing.**

This figure presents a prompt that elicits the LLMs' responses to a question that Tversky and Kahneman (1981) design for documenting narrow framing. The original question from Tversky and Kahneman (1981) writes "20 minutes" for both Scenario A and Scenario B; our paper adjusts it to "5 minutes" in order to account for inflation between 1981 and 2024 (when we collected our data).

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:
Please consider the following scenario. Choose between the following two options: option A, receiving $100 today, versus option B, receiving $110 tomorrow. 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, choose between the following two options: option A, receiving $100 in 30 days, versus option B, receiving $110 in 31 days. 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. IB.5. Prompt for question 5: Hyperbolic discounting.**

This figure presents a prompt that elicits the LLMs' responses to a question discussed in Frederick, Loewenstein, and O'Donoghue (2002) that documents hyperbolic discounting.

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:
Please consider the following scenario. There are two urns. Urn C contains 50 red balls and 50 black balls. Urn U contains 100 balls, each either red or black, but with unknown proportion of each color. Choose between the following two bets: R1: draw a ball from Urn C, get $20 if red, and R2: draw a ball from Urn U, get $20 if red. Please answer as shown above. Indicate the choice you prefer ("R1" or "R2"), 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 an alternative scenario. Specifically, now choose between the following bets: B1: draw a ball from Urn C, get $20 if black, and B2: draw a ball from Urn U, get $20 if black. Please answer as shown above. Indicate the choice you prefer ("B1" or "B2"), 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. IB.6. Prompt for question 6: Ambiguity aversion.**

This figure presents a prompt that elicits the LLMs' responses to a question that Ellsberg (1961) designs for documenting ambiguity aversion.

Instructions:
Consider the following question and respond according to the template provided. 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
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

Question:
A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys.
Which hospital do you think recorded more such days?
A: The larger hospital
B. The smaller hospital
C: About the same (that is, within 5 percent of each other)

Response Format:
Please answer as shown above. Indicate the choice you prefer ("A", "B" or "C"), 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. IB.7. Prompt for question 7: Sample size neglect.**

This figure presents a prompt that elicits the LLMs' responses to a question that Tversky and Kahneman (1974) design for documenting sample size neglect.

Instructions:
Consider the following question and respond according to the template provided. 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
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

Question:
When they turn 18, American males must register for the draft at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 9 inches.
Every day for one year, 25 men registered at post office A and 100 men registered at post office B. At the end of each day, a clerk at each post office computed and recorded the average height of the men who had registered there that day.
Which would you expect to be true (choose one)?
A: The number of days on which the average height was 6 feet or more was greater for post office A than for post office B.
B: The number of days on which the average height was 6 feet or more was greater for post office B than for post office A.
C: There is no reason to expect that the number of days on which the average height was 6 feet or more was greater for one post office than for the other.

Response Format:
Please answer as shown above. Indicate the choice you prefer ("A", "B" or "C"), 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. IB.8. Prompt for question 8: Sample size neglect.**

This figure presents a prompt that elicits the LLMs' responses to a question that Well, Pollatsek, and Boyce (1990) design for documenting sample size neglect.

Instructions:
Consider the following question and respond according to the template provided. 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
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

Question:
You are presented with two covered urns. Both of them contain a mixture of red and green beads. The number of beads is different in the two urns: the small one contains 10 beads altogether, and the large one contains 100 beads altogether. However, the percentage of red and green beads is the same in both urns. The sampling will proceed as follows: You draw a bead blindly from the urn, note its color, and replace it. You mix, draw blindly again, and note down the color again. This goes on to a total of 9 draws from the small urn, or 15 draws from the large urn.
In which case do you think your chances for guessing the majority color are better (choose one)?
A: The small urn that contains 10 beads.
B: The large urn that contains 100 beads.

Response Format:
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. IB.9. Prompt for question 9: Sample size neglect.**

This figure presents a prompt that elicits the LLMs' responses to a question that Bar-Hillel (1979) designs for documenting sample size neglect.

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": {
"Probability": float,
"Confidence": float,
"Explanation": string,
"Reasoning": string
},
"Scenario B": {
"Probability": float,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
}
```

Scenario A:
Please consider the following scenario. Consider the following description of Jack:
"Jack is a 45 year old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles."
Please note that the above description was randomly drawn from a set of 100 descriptions consisting of 70 engineers and 30 lawyers. Given this description, what is the probability that Jack is one of the 70 engineers in the sample of 100?
Please answer as shown above. Indicate the probability that Jack is an engineer (a number between 0 and 1), 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 description of Jack:
"Jack is a 45 year old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles."
Please note that the above description was randomly drawn from a set of 100 descriptions consisting of 30 engineers and 70 lawyers. Given this description, what is the probability that Jack is one of the 30 engineers in the sample of 100?
Please answer as shown above. Indicate the probability that Jack is an engineer (a number between 0 and 1), 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. IB.10. Prompt for question 10: Base rate neglect.**

This figure presents a prompt that elicits the LLMs' responses to a question that Kahneman and Tversky (1973) design for documenting base rate neglect.

Instructions:
Consider the following scenario and respond according to the template provided. 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
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

Scenario:
Please consider the following scenario. Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
Which of the two statements is more probable?
A: Linda is a bank teller.
B. Linda is a bank teller and is active in the feminist movement.
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. IB.11. Prompt for question 11: Conjunction fallacy.**

This figure presents a prompt that elicits the LLMs' responses to a question that Tversky and Kahneman (1983) design for documenting conjunction fallacy.

Instructions:
Consider the following scenario and respond according to the template provided. 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
{
"Probability": float,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

Scenario:
Please consider the following scenario. Imagine you simulate the random outcome of tossing an unbiased coin 150 times in succession. Suppose the last coin toss gave you a head. What is the probability of getting a tail from the next coin toss?
Please answer as shown above. Indicate the probability of getting a tail from the next coin toss (a number between 0 and 1), 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. IB.12. Prompt for question 12: Gambler's fallacy.**

This figure presents a prompt that elicits the LLMs' responses to a question discussed in Rabin (2002) that documents gambler's fallacy.

Instructions:
Consider the following scenario and respond according to the template provided. 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
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

Scenario:
Please consider the following scenario. You are shown four cards, marked E, K, 4 and 7. Each card has a letter on one side and a number on the other. You are given the following rule: Every card with a vowel on one side has an even number on the other side. Which cards must you turn over to test whether the rule is true or false?
Please answer as shown above. Indicate the cards you turn over to test whether the rule is true or false (one or multiple from E, K, 4 and 7), 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. IB.13. Prompt for question 13: Confirmation bias.**

This figure presents a prompt that elicits the LLMs' responses to a question discussed in Wason and Johnson-Laird (1972) that documents confirmation bias.

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": {
"Direction": string,
"Estimate": float,
"Confidence": float,
"Explanation": string,
"Reasoning": string
},
"Scenario B": {
"Direction": string,
"Estimate": float,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
}
```
```

Scenario A:
Please consider the following scenario. Suppose your objective is to estimate the percentage of African countries in the United Nations. For estimating this quantity, consider that a number between 0 and 100 is drawn randomly by spinning a wheel of fortune. Suppose the number drawn is 10.
Please first indicate whether this number of 10 is higher or lower than your estimate on the percentage of African countries in the United Nations. Please then provide your estimate by moving upward or downward from this number of 10.
Please answer as shown above. Indicate the direction ("higher" if 10 is higher than your estimate, and "lower" if 10 is lower than your estimate), your estimate by moving upward or downward from the randomly drawn number of 10 (a number between 0 an 100), 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, suppose your objective is to estimate the percentage of African countries in the United Nations. For estimating this quantity, consider that a number between 0 and 100 is drawn randomly by spinning a wheel of fortune. Suppose the number drawn is 65.
Please first indicate whether this number of 65 is higher or lower than your estimate on the percentage of African countries in the United Nations. Please then provide your estimate by moving upward or downward from this number of 65.
Please answer as shown above. Indicate the direction ("higher" if 65 is higher than your estimate, and "lower" if 65 is lower than your estimate), your estimate by moving upward or downward from the randomly drawn number of 65 (a number between 0 an 100), 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. IB.14. Prompt for question 14: Anchoring.**

This figure presents a prompt that elicits the LLMs' responses to a question that Tversky and Kahneman (1974) design for documenting anchoring.

Instructions:
Consider the following questions and respond according to the template provided. 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
{
{"Questions": [
{"Question number": 1, "Lower bound": float, "Upper bound": float},
{"Question number": 2, "Lower bound": float, "Upper bound": float},
...
],
"Reasoning": string
}
```
```

Questions:
For the following series of questions with clear-cut numerical answers, please provide 90% confidence intervals. Such an interval has a lower and an upper bound such that you are 90% sure that the correct answer lies in this interval. Note that, if your intervals are too wide, the correct answer will fall in your interval more than 90% of the time while, if your intervals are too narrow, the correct answer will fall in your interval less than 90% of the time.

1. World population total growth between 1990 and 2000 (in percentage terms)
2. Year in which Newton discovered universal gravitation
3. Number of nations in OPEC
4. Number of medals that Greece won at the first Olympic Summer Experiments in 1896
5. Year in which Bell patented the telephone
6. Percentage of total area in world covered by water
7. Height of Sears Tower (now known as the Willis Tower) in Chicago (in feet) including the highest antenna on top of the building
8. Number of nations in NATO
9. Age of sun in billions of years
10. Number of joints in human body
11. GDP per capita in France (in thousands of $US) in 2000
12. Current number of member states in the United Nations General Assembly
13. Year in which Mozart wrote his first symphony
14. Gestation (conception to birth) period of an Asian elephant (in days)
15. Elevation (in feet above sea level) of Mt. Everest
16. Number of babies born in world in 2001 (per 1000 people)
17. World-wide life expectancy at birth in 2001
18. Land area in the world (in millions of sq mile as of 2017)
19. Greatest depth (in feet) of the Pacific Ocean
20. Number of calories in 8-ounce russet potato (flesh and skin) according to United States Department of Agriculture

Response Format:
Please answer as shown above. For each question, write the answers as question number (number between 1 and 20), lower bound (a precise number), upper bound (a precise number). For the answers you just provided, please also provide 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. IB.15. Prompt for question 15: Overconfidence - overprecision.**

This figure presents a prompt that elicits the LLMs' responses to a set of general knowledge questions; the questions are adapted from Appendix C of Deaves, Lüders, and Luo (2009) and used to document overprecision.

Instructions:
Consider the following questions and respond according to the template provided. 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
{
{"Questions":  [
{"Question number":  1, "Choice":  string},
{"Question number":  2, "Choice":  string},
...]
,
"Reasoning":  string,
"Accuracy":  int
}
```
```

Questions:
Here are ten questions about investment:
1 - If you buy a company's stock
A. You own a part of the company
B. You have lent money to the company
C. You are liable for the company's debts
D. The company will return your original investment to you with interest

2 - If you buy a company's bond
A. You own a part of the company
B. You have lent money to the company
C. You are liable for the company's debts
D. You can vote on shareholder resolutions

3 - If a company files for bankruptcy, which of the following securities is most at risk of becoming virtually worthless?
A. The company's preferred stock
B. The company's common stock
C. The company's bonds

4 - In general, investments that are riskier tend to provide higher returns over time than investments with less risk.
A. True
B. False

5 - Over the 30 years ending in December 2019 in the US, the best average returns have been generated by:
A. Stocks
B. Bonds
C. CDs
D. Money market accounts
E. Precious metals

6 - What has been the approximate average annual return of the S&P 500 stock index for the 50 years ending in December 2019 (not adjusted for inflation)?
A. –10%
B. –5%
C. +5%
D. +10%
E. +15%
F. +20%

7 - Which of the following best explains the distinction between nominal returns and real returns?
A. Nominal returns are pre-tax returns; real returns are after-tax returns
B. Nominal returns are what an investment is expected to earn; real returns are what an investment actually earns
C. Nominal returns are not adjusted for inflation; real returns are adjusted for inflation
D. Nominal returns are not adjusted for fees and expenses; real returns are adjusted for fees and expenses

8 - Which of the following best explains why many municipal bonds pay lower yields than other government bonds?
A. Municipal bonds are lower risk
B. There is a greater demand for municipal bonds
C. Municipal bonds can be tax-free

9 - You invest $500 to buy $1,000 worth of stock on margin. The value of the stock drops by 50%. You sell it. Approximately how much of your original $500 investment are you left with in the end?
A. $500
B. $250
C. $0

10 - Which is the best definition of selling short?
A. Selling shares of a stock shortly after buying it
B. Selling shares of a stock before it has reached its peak
C. Selling shares of a stock at a loss
D. Selling borrowed shares of a stock

Response Format:
Please answer as shown above. For each question, write your answers as question number (number between 1 and 10) and your choice (one of either "A", "B", "C", "D", "E", or "F") . For the answers you just provided, please also provide 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) and provide an estimate of the accuracy of your answers reflecting how many questions you believe you answered correctly (an integer between 0 and 10).

**Fig. IB.16. Prompt for question 16: Overconfidence - overestimation.**

This figure presents a prompt that elicits the LLMs' responses to a set of ten questions about investment and their estimate of the accuracy of their responses; the prompt follows the procedure discussed in Moore and Healy (2008) to document overestimation and the ten questions are based on the "investing knowledge quiz" designed by the Financial Industry Regulatory Authority.