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*

(Preliminary and incomplete; updates to follow.)

ABSTRACT

Do generative AI models as epitomized and popularized by large language models (LLMs) exhibit systematic behavioral biases, especially 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 parameter scale. We document systematic behavioral biases exhibited by LLMs. For experiments concerning the psychology of beliefs, LLM responses become more rational as the models become more advanced or larger; for experiments concerning the psychology of preferences, even the most advanced large-scale models frequently generate irrational and human-like responses. Further exploring various methods for correcting these behavioral biases reveals that prompting LLMs to behave as rational investors who make 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, Siew Hong Teoh, 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 exceptional 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 whether LLMs exhibit specific biases in decision making, with a focus on the behavior of ChatGPT.[1] Our paper not only adds to these studies but also aims to establish benchmark results for the new field of behavioral economics of AI: 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 our paper is to develop a public database of experimental questions for the ongoing evaluation 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 the experimental economics literature, which, compared to the cognitive psychology literature, includes 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 beliefs, the LLMs' answers exhibit a clear pattern: as we progress to more advanced models or those with a larger parameter scale, the responses become 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 1.5 Pro—produce by and large rational answers to belief-based questions.

Second, when asked questions from the cognitive psychology literature that document biases in preferences, the LLMs' answers exhibit the *opposite* pattern: for models that are more advanced or with a larger parameter scale, the responses become increasingly human-like and they are not rational 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,

[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.

gives human-like answers only to one out of the six questions.

Third, we observe substantial heterogeneity in LLM responses when comparing responses across the four different families of LLMs. For belief-based questions, responses from Meta Llama are less rational and more human-like compared to those from ChatGPT, while the responses from Anthropic Claude or Google Gemini are by and large similar to those from ChatGPT. For preference-based questions, responses from Gemini are less rational and more human-like compared to those from GPT, while the responses from Claude or Llama are by and large 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 three methods for correcting the observed behavioral biases. Among these methods, one is effective while the other two 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 two methods involve combining the brief sentence that primes a LLM to be a rational investor with the provision of additional bias-reducing information.

[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.

These two methods are 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 understanding of the underlying mechanisms is beyond the scope of the paper, two conjectures are worth noting. First, why do more advanced or larger-scale models become more human-like when responding to preference-based questions? We speculate 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 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; 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 help 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 is important for addressing 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 remains 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 largely assume that LLMs function as neutral research tools, implicitly treating their responses as unbiased. However, 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 exhibited by LLMs——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. Moreover, understanding the behavior of AI agents is 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, although prompt-based instructions that explicitly require unbiased decisions are successful in reducing these biases. And Ouyang, Yun, and Zheng (2024) study how risk preferences of LLMs in financial settings can be modulated by techniques that are designed to align LLM behavior with human ethical standards. 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 more systematic and comprehensive in several ways. First, prior studies focus on specific cases, i.e., either a single LLM—typically a version of GPT—or isolated aspects of LLM behavior, such as a concrete rationality measure or a specific type of behavioral bias. By contrast, our paper systematically documents 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 literature, and we cover both experimental questions that study the psychology of preferences and experimental questions that study the psychology of beliefs.[6] Our exploration of debiasing methods is also more comprehensive: we compare different methods and propose new ones, with relevance for practical 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] Finally, we are among the first to advocate treating behavioral economics of AI as a new research field and treating GenAI agents as a new species.

The rest of the paper proceeds as follows. Section 2 discusses the experimental design. Section 3 presents our main results on LLMs' responses to preference-based and belief-based questions. Section 4 explore the methods that aim at 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. Section 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

[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 needed a longer data sample that allows us 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.

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.

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 an example, 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 payoff, 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 below 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 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 ρ measures the persistence of the process and ϵ_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 μ, a given ρ in the range [0,1], and an ϵ_t that is an i.i.d. random shock."

For each of the three experiments and for a wide range of values of ρ, we compare ρ 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. In particular, we note important differences along the three following 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.[8,9]

In terms of model architecture, we further 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.[10] 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.[11,12]

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. RLHF was initially implemented by Ouyang et al. (2022) to fine-tune ChatGPT using human feedback. Today, each family of LLMs uses its proprietary RLHF; for ex-

[8]For more details, please visit Meta Llama 3 characteristics on Meta Llama at: https://ai.meta.com/blog/meta-llama-3/.

[9]For the other three families, although the exact training data are not often disclosed publicly, 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 GPT 3.5. While the exact number of tokens used to train GPT-4 is not known, 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/.

[10]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 the development of larger models. 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.

[11]Two commonly used compression techniques are: quantization, which reduces parameter precision, and pruning, which removes less important connections from the neural network.

[12]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).

ample, 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 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 the 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 specific experimental scenarios; in Fig. 1, this part starts from "Instructions" and ends with "completely separate from the other." 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 " ‘‘‘". [13] 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)." At the end of each question, 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 four different parts: choice, confidence, explanation, and reasoning. Here, "choice" refers

[13] 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.

to an explicit choice made by the LLM—for example, whether the LLM accepts or turns down a risky gamble.[14] "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" corresponds to reasoning that is based more on analytical thinking and calculations.[15] 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."[16]

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 design 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.[17] For the remaining LLMs that support graphical inputs, we follow platform-specific guidelines when uploading figures.[18] Second, the

[14]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.

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

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

[17]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.

[18]Google Gemini directly processes figures that are uploaded as .jpg files. ChatGPT and Anthropic Claude, however, require first encoding a binary image input into bytes and then converting the byte output into a regular UTF-8 string format.

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 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

In this section, we 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 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 parameter 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.[19] Note that setting the temperature parameter to zero results in deterministic outputs, while higher values increase the randomness in LLM responses.[20]

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 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). Table 3 provides the

[19] 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].

[20] Specifically, for the iterative process of generating each word (token) in a response, the LLM first produces a probability distribution over all possible tokens in its dictionary and then draws the next token from this distribution. The temperature parameter reshapes the distribution: 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.

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 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 a total of twelve models. These include three versions 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 parameter 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 category.

[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 by and large similar to those from GPT models.

Again, 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 by and large 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. As such, we control for LLM family fixed effects in our subsequent analyses of heterogeneity across model generations and parameter scales.

3.2.2. Heterogeneity across model generations and parameter scales

We next examine variations in LLM responses across model generations and parameter scales. The evolutions of model generation and parameter scale capture the 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 similar scale. To study the effect of parameter 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 parameter 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 correspond to the underlying data reported in Table 3 and offer a compact view of cross-model variations. For example, Claude 3 Opus does not produce predominantly rational responses for any preference-based question, while Claude 3 Haiku produces predominantly rational responses for three out of six preference-based questions.

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 questions, the left panel of Fig. 3 shows that, as the LLMs become more advanced or larger in parameter scale, the number of questions that receive predominantly rational responses tends to decrease, while the number of questions that receive 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 parameter scales, we estimate a series of probit regressions. In particular, we conduct two analyses. First, to study the effect of model generation on LLM responses, we restrict 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, γ_f captures the LLM family fixed effects.

Second, to study the effect of parameter 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 parameter 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 parameter 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 parameter 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 parameter 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 parameter 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 parameter scales. As we progress to more advanced or larger-scale models, the LLM 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 the behavior of LLMs.

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; these experiments are described in Section 2.1 and are labeled "Experiment 1," "Experiment 2," and "Experiment 3." For each of the experiments, we simulate the autoregressive process:

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

specified in (1), by setting μ, the constant term, to 0 and σ, the standard deviation of ϵ_t, to 20; these parameter values are taken from Afrouzi et al. (2023). For ρ, 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 different paths. As such, each experiment has a total of 600 simulated 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. We begin by randomly selecting one of the 600 simulated paths. In the first round, we present each LLM with a figure that displays the first 40 realizations of x_t from this simulated path. Then, a prompt requests the LLM to provide its forecasts for the next two values: x_{41} and x_{42}. The model's response is recorded and used to establish the beginning of a conversation history. In the second round, we first update the conversation history by appending the previous figure, prompt, and LLM response. We then present a new figure that extends the observed sequence to x_{41} and prompt the LLM to forecast the next two values, x_{42} and x_{43}. We record the LLM's response and append 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 the LLM's forecasts. Specifically, for each LLM i, each value of ρ, and each forecast horizon s of 1, 2, or 5, we collect 500 forecasts and 500 realizations of x_t.

We then estimate the perceived persistence $\hat{\rho}$ by running the following regression:

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

where $F_{it}x_{t+s}$ represents model i's forecast of x_{t+s} at time s, and x_t is the actual realization of x at time t.

Fig. 4 presents our estimates. The top panel displays the estimated $\hat{\rho}$ values for three baseline models: ChatGPT-4, Claude 3 Opus, and Gemini 1.5 Pro. The bottom panel shows the same estimates for the three smaller-scale models: ChatGPT-4o, Claude 3 Haiku, and Gemini 1.5 Flash.[21] Each estimate includes a 95% confidence interval. The results are compared to the 45-degree line, which represents the implied persistence under Full Information Rational Expectations (FIRE).

[Place Fig. 4 about here]

The results show that baseline models behave rationally, while smaller-scale models exhibit human-like biases. The larger models produce similar results, which we classify as rational behavior based on two evidence. First, we notice that baseline models do not display persistent over- or underreaction. For small values of ρ (≤ 0.2), the estimates of $\hat{\rho}$ show overreaction implying that LLMs perceive the autoregressive process as more persistent than it actually is. For large values of ρ ($\rho \geq 0.4$), the models exhibit underreaction, meaning they underestimate the actual persistence of the autoregressive process. This pattern differs from human behavior, where overreaction occurs consistently across all values of ρ. Second, the extent of over- and underreaction remains stable across values of ρ for baseline models. Unlike human participants—who display stronger overreaction for smaller values of ρ—the baseline LLMs exhibit a relatively constant level of deviation from the FIRE benchmark. Small-scale models exhibit human-like overreaction. These models consistently overreact, similar to human participants. Moreover, their degree of overreaction is larger for smaller values of ρ, aligning with the cognitive biases in human forecasting behavior documented by Afrouzi et al. (2023).

Overall, our findings are consistent with those from the cognitive psychology literature on beliefs: larger models exhibit significantly more rational behavior, while smaller-scale models display

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

human-like biases. This result reinforces the broader pattern observed in our experiments—LLMs with larger architectures tend to align more closely with normative rationality, whereas smaller models retain cognitive distortions characteristic of human decision-making.

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 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.

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

Table 6 suggests that instructing the LLMs to behave as rational investors is effective in reducing biases; that is, 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. The specific four-step procedure is given by:

"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 probability 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

[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 the prospect theory-related questions.

human-like responses by about 18%. Taken together, these findings suggest that provision of more information—even if such information is genuinely useful for decision making—is not always useful in correcting LLM biases: information overload might hinder a LLM's ability to provide rational responses.

5. Conclusion

Artificial intelligence, especially generative AI epitomized by LLMs, has become increasingly important in social and economic activities. In this paper, we systematically 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.

Overall, LLMs exhibit systematic behavioral biases. For experimental questions that study human beliefs, the LLMs' responses become more rational as we move towards more advanced models or models with a larger parameter scale; here, we examine belief-based questions from both the psychology and experimental economics literatures. For questions that study human preferences, however, even the most advanced large-scale LLMs frequently generate responses that are irrational and human-like. Moreover, we observe significant heterogeneities in the LLM responses across the four families of LLMs.

We also explore role priming methods that 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 to behave as real-world retail investors leads to less rational responses when the LLMs answer preference-based questions. Finally, we show that 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 necessarily 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.

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.

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. 38th AAAI Conference on Artificial Intelligence 38, 17960–17967.

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.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., et al, 2022. Training Language Models to Follow Instructions with Human Feedback. Working Paper.

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

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. 34th Conference on Neural Information Processing Systems pp. 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., 1981. The Framing of Decisions and the Psychology of Choice. Science 211, 453–458.

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.

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.

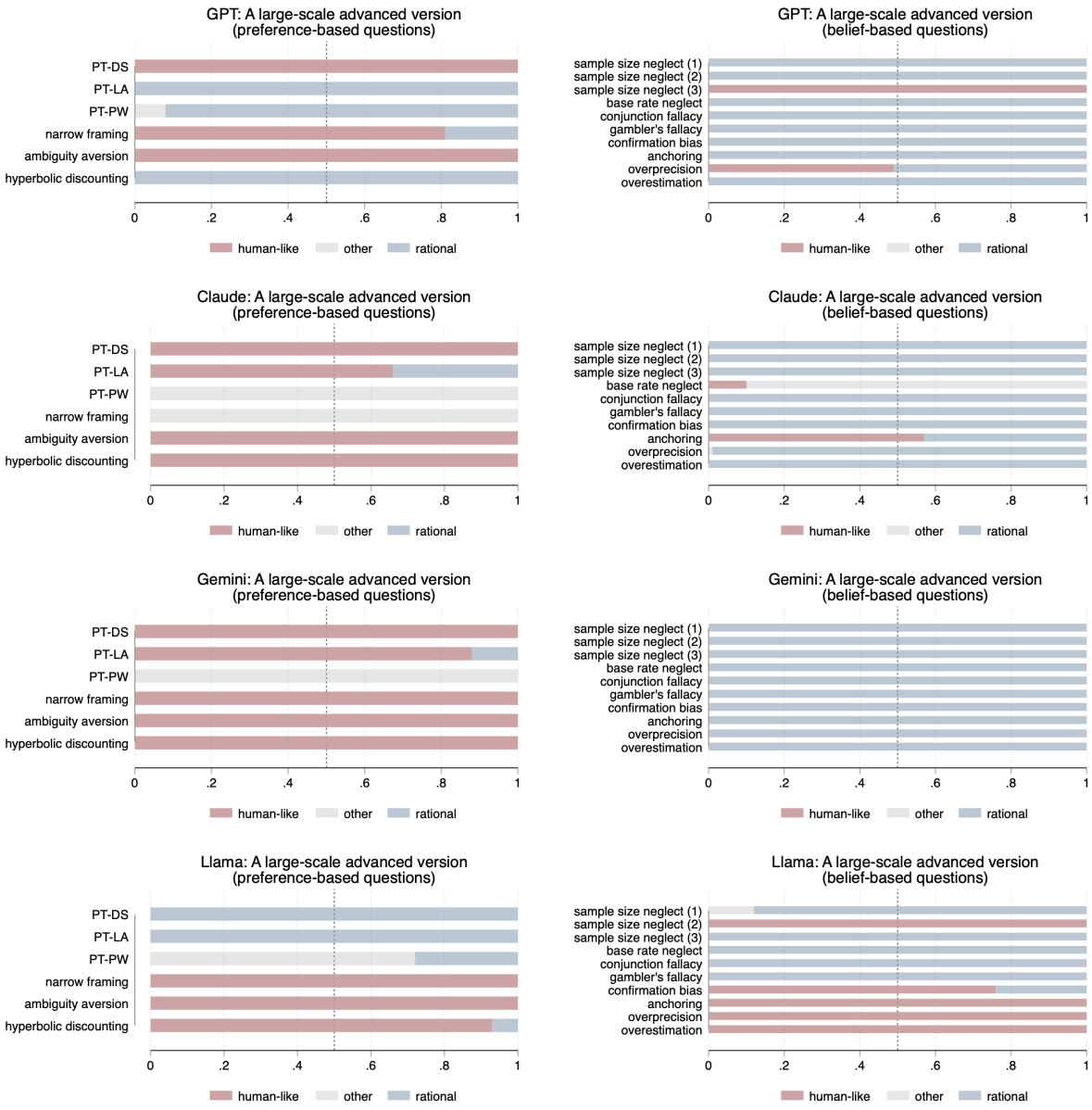


Fig. 2. Proportion of LLM responses: Advanced large-scale models.

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.

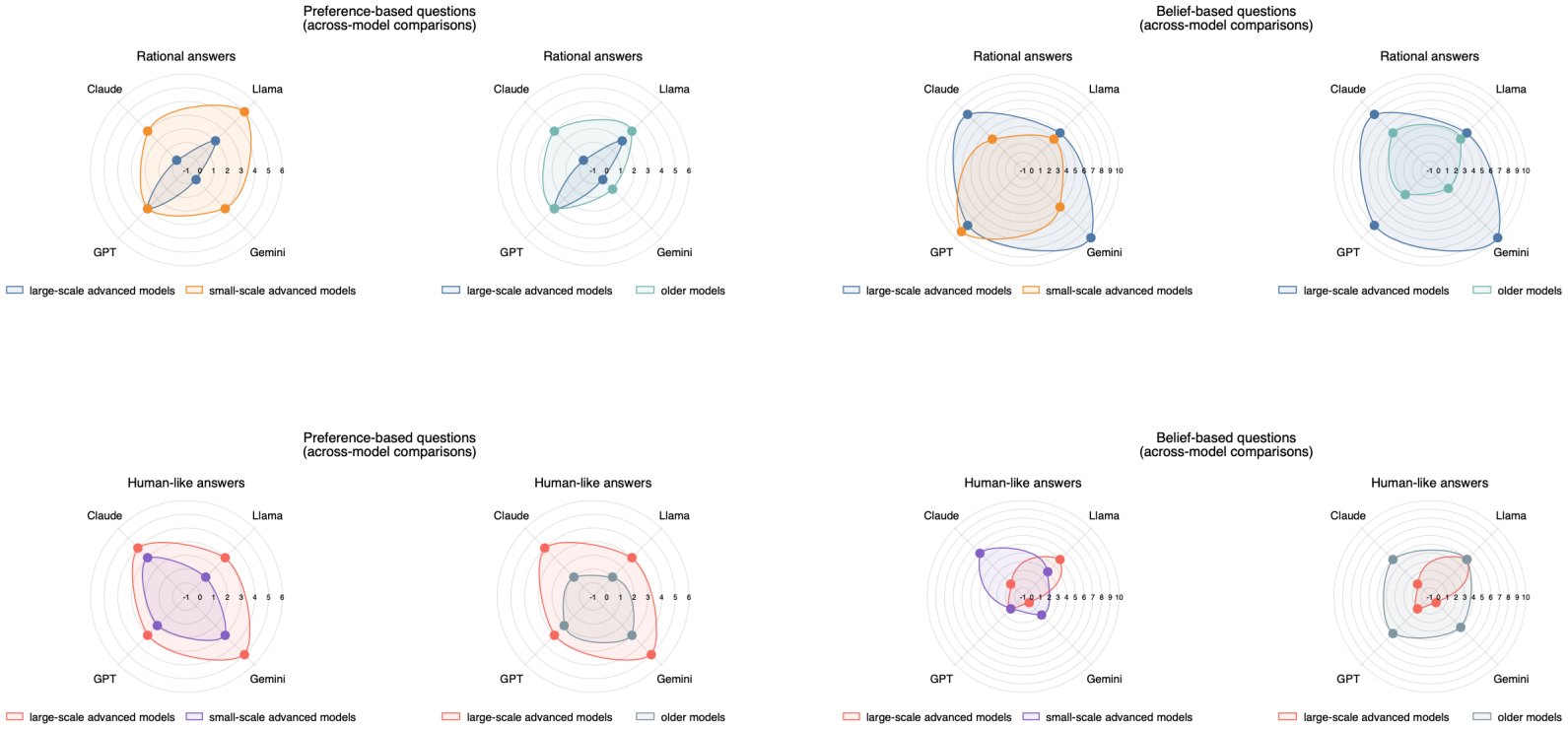
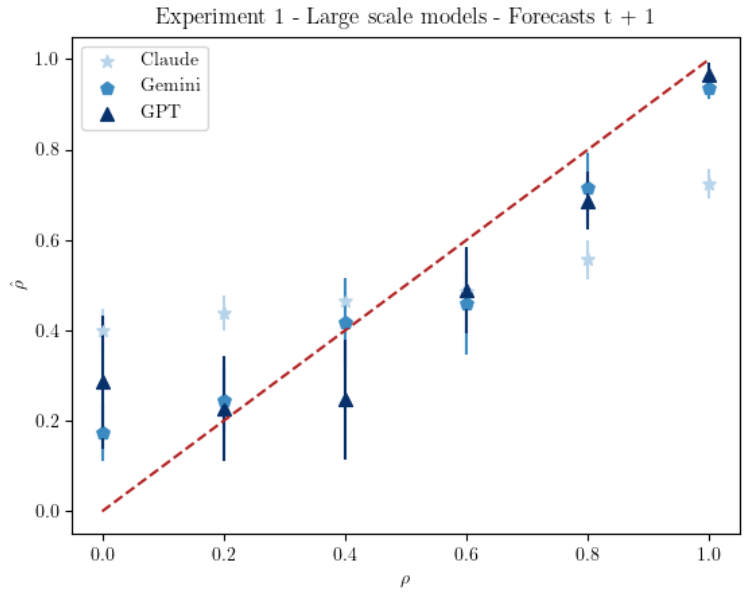


Fig. 3. Heterogeneity in LLM responses across model generations and parameter scales.

This figure presents radar charts that compares 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 panels) and belief-based questions (right panels). Comparisons are made between advanced large-scale models and advanced smaller-scale models, and between large-scale advanced models and older models.

Panel A



Panel B

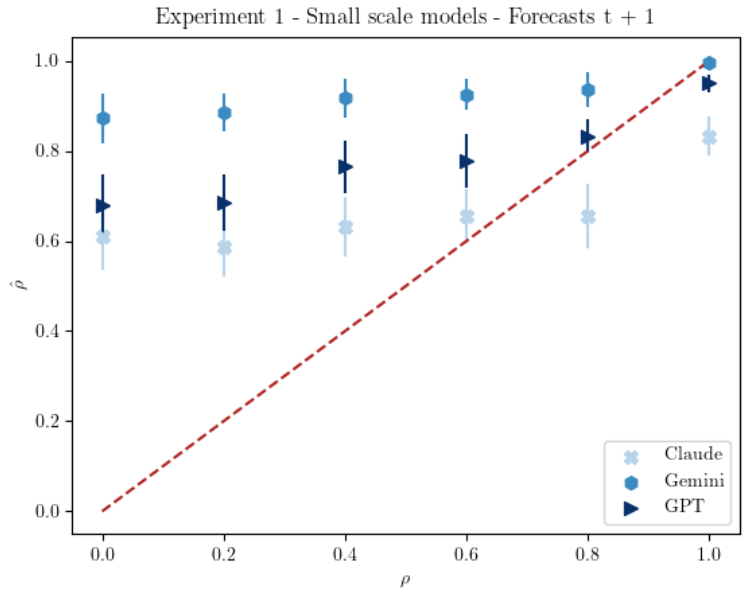
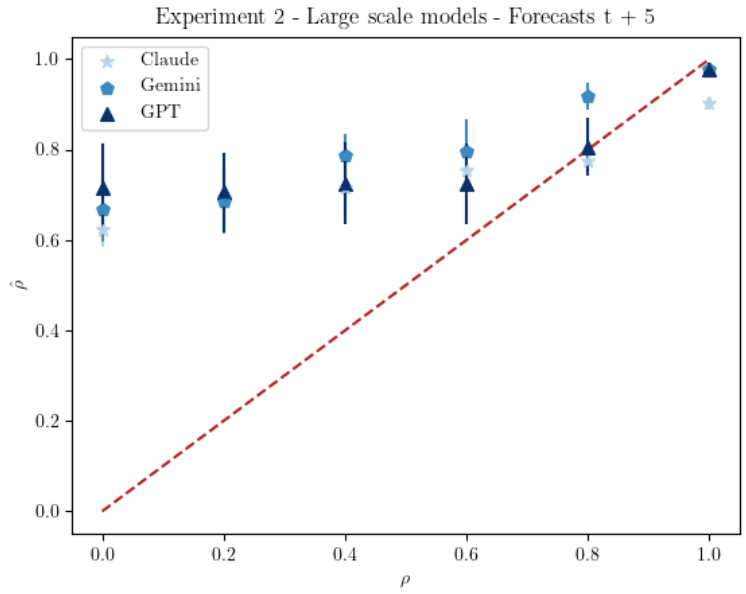


Fig. 4. LLM responses to questions from Experiment 1 of Afrouzi et al. (2023).

The figure plots forecast-implied persistence (y-axis) and the actual AR(1) persistence (x-axis) for questions based on Experiment 1 of Afrouzi et al. (2023). For each level of AR(1) persistence ρ, we estimate the implied persistence $\hat{\rho}$ from $F_{it}x_{t+s} = c_s + (\hat{\rho})^s x_t + u_{s,it}$. The red line is the 45-degree line, and corresponds to the implied persistence under Full Information Rational Expectations (FIRE). The vertical bars show the 95% confidence interval of the point estimates. Top panel reports the results for three large-scale advanced LLMs that allow image input: GPT-4, Claude 3 Opus, and Gemini 1.5 Pro; bottom panel reports the results for three small-scale advanced LLMs that allow image input: GPT-4o, Claude 3 Haiku, and Gemini 1.5 Flash.

Panel A



Panel B

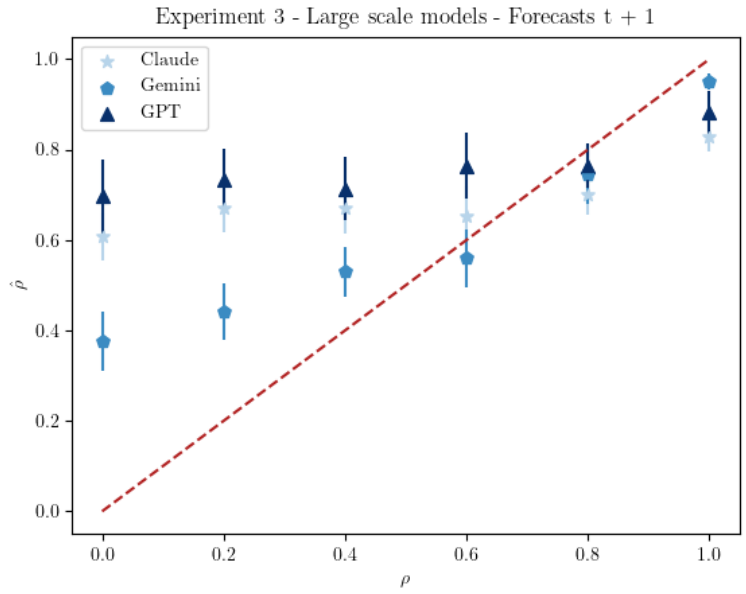


Fig. 5. LLM responses to questions from Experiments 2 and 3 of Afrouzi et al. (2023).

The figure plots forecast-implied persistence (y-axis) and the actual AR(1) persistence (x-axis) for questions based on Experiments 2 and 3 of Afrouzi et al. (2023). For each level of AR(1) persistence ρ, we estimate the implied persistence $\hat{\rho}$ from $F_{it}x_{t+s} = c_s + (\hat{\rho})^s x_t + u_{s,it}$. The red line is the 45-degree line, and corresponds to the implied persistence under Full Information Rational Expectations (FIRE). The vertical bars show the 95% confidence interval of the point estimates. Top panel reports the results for three large-scale advanced LLMs that allow image input: GPT-4, Claude 3 Opus, and Gemini 1.5 Pro; bottom panel reports the results for three small-scale advanced LLMs that allow image input: GPT-4o, Claude 3 Haiku, and Gemini 1.5 Flash.

Table 1. Summary of experimental questions from cognitive psychology.

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
8	sample size neglect
9	sample size neglect
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.

We group twelve LLMs based on the four LLM families that we consider: OpenAI's 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.

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 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

	Claude				GPT				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	0.34	(1.000)	0.66	(0.001)***	1.00	(0.000)***	0.00	(1.000)	0.12	(1.000)	0.88	(0.000)***	1.00	(0.000)***	0.00	(1.000)
PT-PW	0.00	(1.000)	0.00	(1.000)	0.92	(0.000)***	0.00	(1.000)	0.00	(1.000)	0.00	(1.000)	0.28	(1.000)	0.00	(1.000)
narrow framing	0.00	(1.000)	0.00	(1.000)	0.19	(1.000)	0.81	(0.000)***	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	0.00	(1.000)	1.00	(0.000)***	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***	0.07	(1.000)	0.93	(0.000)***

Panel B: Belief-based questions

	Claude				GPT				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)	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)
base rate neglect	0.00	(1.000)	0.10	(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)
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	0.43	(0.933)	0.57	(0.097)*	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***
overprecision	0.99	(0.000)***	0.00	(1.000)	0.51	(0.460)	0.49	(0.618)	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 parameter 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 those 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 probit regressions where the dependent variable is an indicator for whether a LLM response is classified as rational or human-like. Regressions in Columns (1) to (4) are for preference-based questions and those 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 probit regressions where the dependent variable is an indicator for whether a LLM response is classified as rational or human-like. 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

Appendix

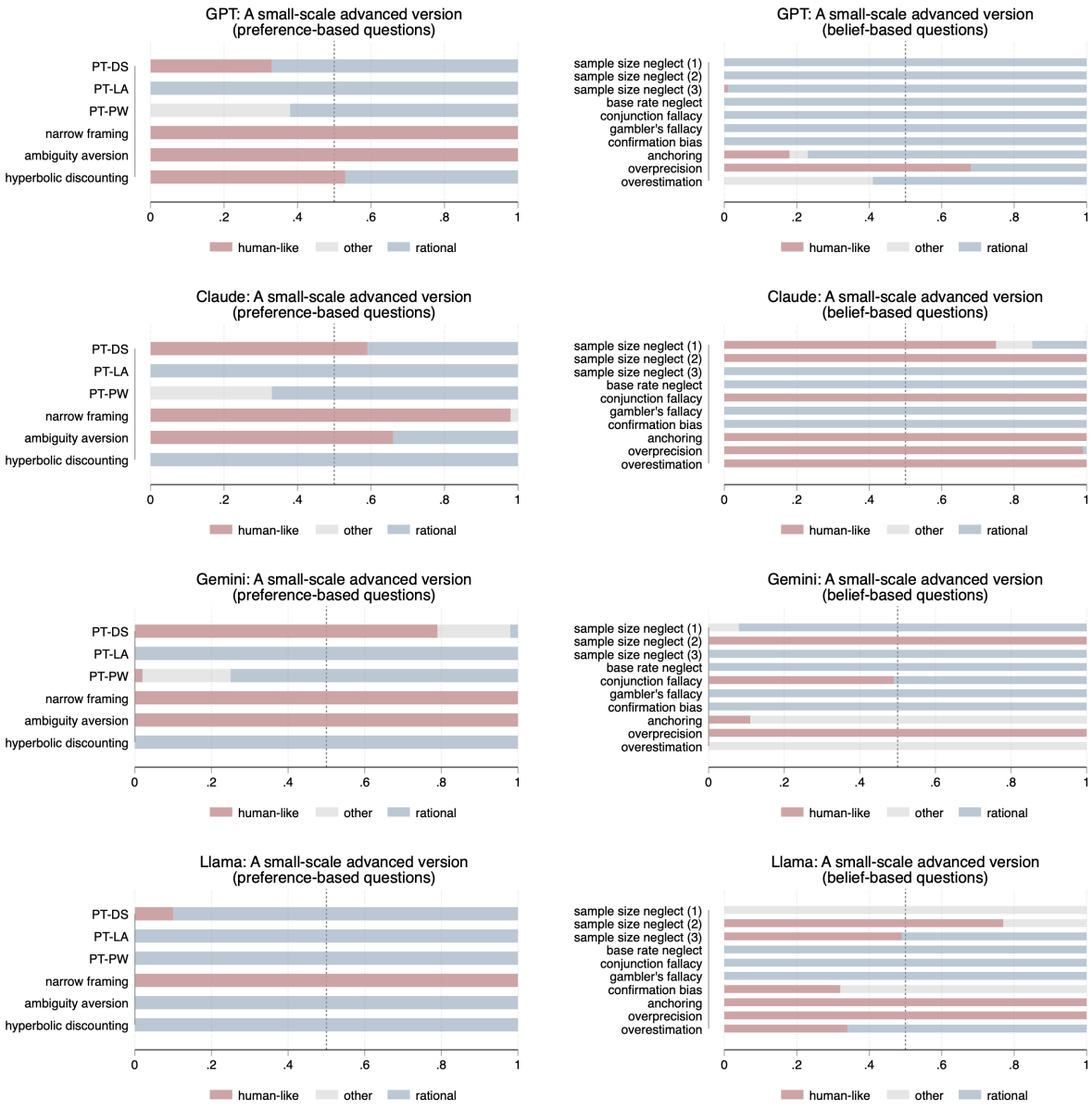


Fig. IA.1. Proportion of LLM responses: Advanced small-scale models.

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.

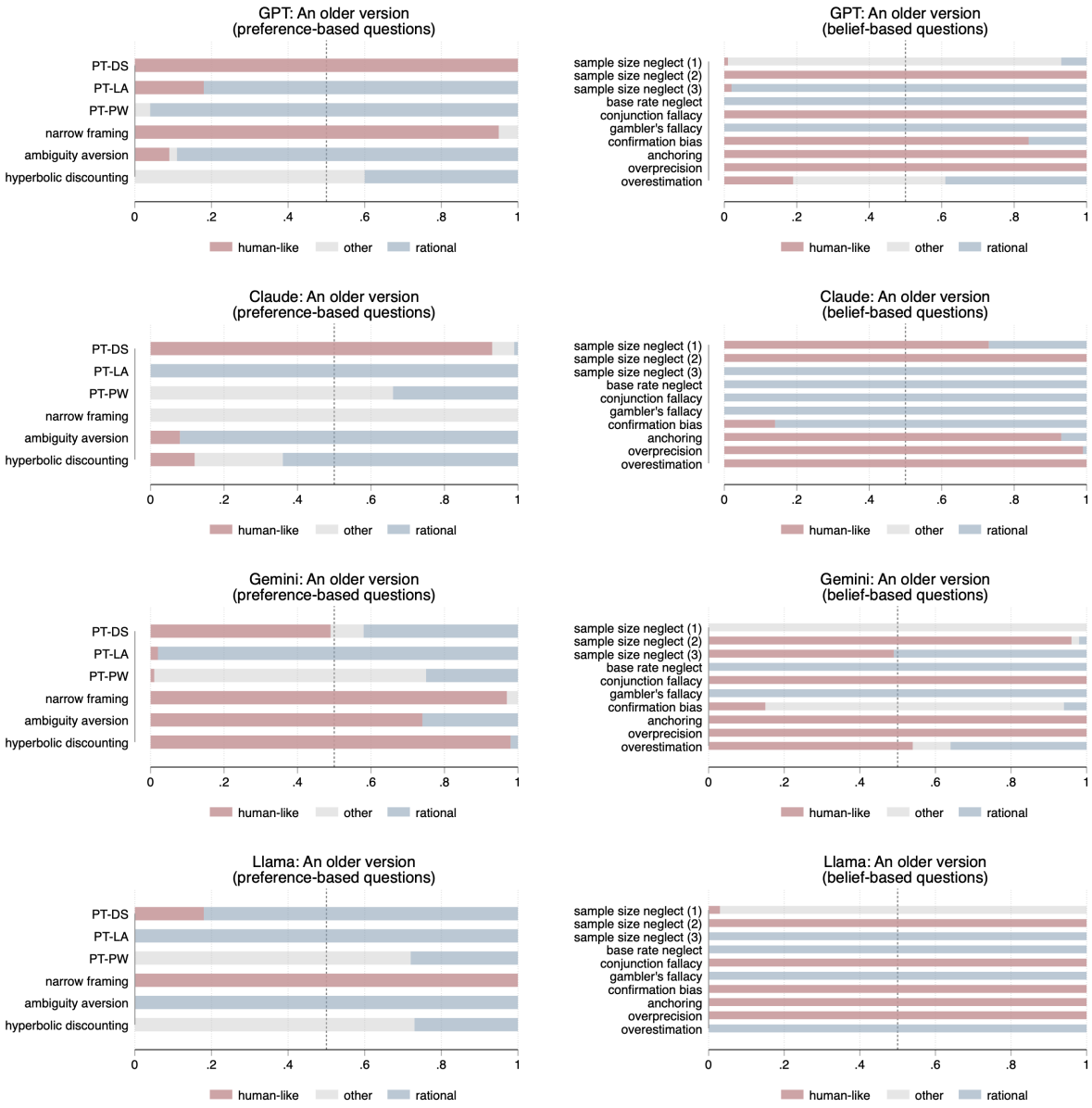


Fig. IA.2. Proportion of LLM responses: Older models.

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.