Behavioral Economics of AI: LLM Biases and Corrections

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

1 Introduction

Artificial intelligence (AI), especially generative large language models (LLMs), is becoming increasingly essential in daily work and general economic activities. 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 et al., 2024]. Researchers are investigating the potential for LLMs to enhance experimentation that studies human behavior [Charness et al., 2023, Korinek, 2023, Bail, 2024]. However, little is known about how AI algorithms and agents behave systematically, especially in economic and financial decisions, let alone whether their behavior closely resembles that of humans. Understanding the "behavioral economics" of AI—potentially a new intelligent life form [Tegmark, 2017]—starting with LLMs is urgent and crucial for assessing and improving the technology's utility, safety, and appropriateness.

Recent studies have started to examine the reliability of LLMs in decision making, with a focus on specific behavior of ChatGPT in both individual decision-making settings [Chen et al., 2023, Ma et al., 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]. Chen et al. [2023] find that GPT-3.5 Turbo exhibits higher economic rationality and lower choice heterogeneity compared to humans. Mei et al. [2024] show GPT-4 exhibits human-like behavioral traits in games. Chen et al. [2025] document that LLMs manifest biased beliefs when forecasting stock returns. Bowen et al. [2025] find racial biases in LLM mortgage underwriting recommendations, mitigatable through prompts. Ouyang et al. [2024] study how risk preferences of LLMs can be modulated by alignment techniques.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Our paper conceptually introduces behavioral economics of AI as a new field through establishing its benchmark results: we conduct the most comprehensive set of experiments to date, originally designed to document human biases, but now applied to investigate the biases of multiple prominent families of LLMs; we systematically compare LLM responses with both rational responses and human responses; and we explore methods for correcting their biases. An important goal is developing a public database of experimental questions for ongoing evaluations of behavioral biases in various LLMs.

We begin by drawing on the cognitive psychology literature, originated by Ellsberg [1961] and Kahneman and Tversky [1973, 1979], using carefully designed experimental questions to assess psychological biases. From this literature, we select comprehensive experiments covering both psychology of preferences and beliefs, ensuring inclusion of those documenting biases of first-order importance in financial markets—prospect theory preferences, overextrapolation, and overconfidence [Barberis, 2018]. We also turn to recent experimental economics studies [Lian et al., 2018, Bose et al., 2022, Afrouzi et al., 2023], adapting tasks more closely tied to economic and financial settings. With experimental questions at hand, we collect responses through APIs from four prominent LLM families: OpenAI's ChatGPT, Anthropic Claude, Google Gemini, and Meta Llama, examining both advanced versus older versions and large versus smaller scales.

2 Experimental Design

2.1 Selection of Experimental Questions

Traditional theories in economics and finance posit that economic agents make rational decisions with two components: rational preferences (Expected Utility framework by Von Neumann and Morgenstern [1944]) and rational beliefs (Bayesian updating). However, decades of cognitive psychology research cast doubt on such theories through carefully designed experimental questions documenting systematic deviations. To illustrate, Kahneman and Tversky [1979] posed: "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. Most participants choose B. Then asked separately: "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)." Most choose C. Yet A equals C and B equals D in monetary payoffs, violating Expected Utility.

Table 1 summarizes sixteen experimental questions we study. Each LLM response is classified as: rational (derived from Expected Utility and Bayes' law), human-like (irrational but matching typical human responses), or other (neither). The table covers questions documenting prospect theory preferences (Questions 1-3: diminishing sensitivity, loss aversion, probability weighting), overextrapolation (Questions 7-10: sample size neglect, hot hand fallacy, law of small numbers, base rate neglect), and overconfidence (Questions 15-16: overprecision, overestimation)—the "big three" biases of first-order importance in financial markets according to Barberis [2018].

Additionally, following Afrouzi et al. [2023], we ask LLMs to observe past realizations of random variable x_t and forecast future realizations, where time-series evolution follows: $x_t = \mu + \rho x_{t-1} + \epsilon_t$, with ρ measuring persistence and ϵ_t being i.i.d. Gaussian. We conduct three experiments: (1) baseline with knowledge that x_t follows a "stable random process," observing 40 past realizations then making five rounds of forecasts for x_{t+1} and x_{t+2}; (2) variant forecasting x_{t+1} and x_{t+5} for longer horizons; (3) variant with detailed knowledge that evolution follows "a fixed and stationary AR(1) process: $x_t = \mu + \rho x_{t-1} + \epsilon_t$, with given μ, given ρ in [0,1], and i.i.d. random shock ϵ_t." For each experiment and $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, we compare true ρ with perceived $\hat{\rho}$ implied by LLM forecasts to document belief biases.

2.2 Selection of LLMs and Prompt Design

We select twelve LLMs from four prominent GPT families (Table 2). For each family: benchmark model (most recent/best-performing), smaller-scale version, and predecessor. For ChatGPT: GPT-4 (benchmark), GPT-4o (smaller), GPT-3.5 Turbo (predecessor). For Anthropic Claude: Claude 3 Opus (benchmark, 200B parameters), Claude 3 Haiku (smaller), Claude 2 (predecessor, 100K token context). For Google Gemini: Gemini 1.5 Pro (benchmark), Gemini 1.5 Flash (distilled smaller),

Gemini 1.0 Pro (predecessor). For Meta Llama: Llama 3 70B (benchmark), Llama 3 8B (smaller), Llama 2 70B (predecessor).

These models differ in: training data (Llama 3: 15 trillion tokens vs. Llama 2: 1.8 trillion; GPT-3.5: 500B tokens vs. GPT-4: 13 trillion unofficial), architecture (single-transformer vs. mixture-of-experts; GPT-4 reportedly 8×220B parameters totaling >1 trillion), and RLHF implementations (Claude's Constitutional AI for helpfulness/harmlessness/honesty [Bai et al., 2022]).

Our prompts have three parts (Figure 1): (1) general instructions ("Instructions: You will be presented with a series of experimental scenarios...treat each scenario as completely separate from the other"); (2) JSON response format ("The output should be formatted as a markdown snippet containing a JSON object literal within a code block"); (3) experimental questions with scenario-specific instructions. Responses contain: choice (explicit selection), confidence (0-1 score), explanation (brief justification), reasoning (A: intuitive thinking, B: analytical calculations). For within-subject designs, we combine multiple questions per API call, treating each call as an individual participant. Temperature parameter set to 0.5 (recommended), with top-k=50, top-p=0.9.

For Afrouzi experiments requiring graphical inputs (figures displaying past realizations), six models lacking this capability are excluded (Llama models, GPT-3.5 Turbo, Claude 2, Gemini 1.0 Pro). We encode images as UTF-8 strings for ChatGPT/Claude, direct .jpg for Gemini. Sequential API calls preserve conversation history across forecast rounds.

3 Behavioral Biases of LLMs

3.1 Baseline Results

For each question and model, we collect 100 responses (100 API calls per question-model pair). Figure 2 and Tables 3, 4, 5, and 6 present results for benchmark models (GPT-4, Claude 3 Opus, Gemini 1.5 Pro, Llama 3 70B), categorizing responses as rational (blue), human-like (red), or other (gray).

Two key observations emerge. First, most responses fall into rational or human-like categories; "other" responses are rare (GPT-4: only Question 3 on probability weighting; Claude 3 Opus: Questions 3, 4, 10, 15; Gemini 1.5 Pro: Question 3; Llama 3 70B: Questions 3, 7). Second, preference-based questions (left panel) elicit predominantly human-like responses while belief-based questions (right panel) elicit predominantly rational responses. Binomial tests (null: proportion $\leq 50\%$) confirm patterns with >99% confidence. For preferences: Gemini 1.5 Pro has majority human-like in 5/6 questions, Claude 3 Opus 4/6, GPT-4 and Llama 3 70B 3/6. For beliefs: Gemini 1.5 Pro has majority rational in 10/10 questions, GPT-4 and Claude 3 Opus 8/10, Llama 3 70B 5/10.

3.2 Heterogeneity in LLM Responses

Across Families: Using probit regression $\Pr(Y_{iqk} = 1) = \Phi(\alpha + \beta_1 \cdot Claude_i + \beta_2 \cdot Gemini_i + \beta_3 \cdot Llama_i + \epsilon_{iqk})$ where Y_{iqk} indicates rational (or human-like) response for model i, question q, iteration k, Table 7 reports marginal effects. For preferences, Gemini models are 22.9% less likely to produce rational responses (p<0.01) and 16.7% more likely to be human-like (p<0.05) versus GPT baseline. Claude and Llama responses statistically similar to GPT. For beliefs, Llama models are 25.0% less likely rational (p<0.05) and 21.0% more likely human-like (p<0.05) versus GPT. Claude and Gemini statistically similar to GPT.

Across Generations and Scales: Figure 3 displays radar charts showing questions receiving predominantly rational/human-like responses. For studying generation effects, we compare advanced versus older models using $\Pr(Y_{iqk} = 1) = \Phi(\alpha + \beta \cdot Advanced_i + \gamma_f + \epsilon_{iqk})$ with family fixed effects γ_f. For scale effects, we compare large versus small using $\Pr(Y_{iqk} = 1) = \Phi(\alpha + \beta \cdot LargeScale_i + \gamma_f + \epsilon_{iqk})$. Table 8 reports marginal effects. For preferences: advanced models 15.2% less likely rational (p<0.01), 12.4% more likely human-like (p<0.01); large-scale models 13.8% less likely rational (p<0.01), human-like increase insignificant. For beliefs: advanced models 28.3% more likely rational (p<0.01), 24.9% less likely human-like (p<0.01); large-scale models 19.7% more likely rational (p<0.01), 17.3% less likely human-like (p<0.01).

3.3 Responses to Afrouzi Experiments

We estimate perceived persistence $\hat{\rho}$ using: $F_{it}x_{t+s} = c_{is} + (\hat{\rho}_{is})^s x_t + u_{is,t}$ where $F_{it}x_{t+s}$ represents model i's time-t forecast of x_{t+s}. Figure 4 shows for short-term forecasts ($s = 1$), advanced large-scale models (GPT-4, Claude 3 Opus, Gemini 1.5 Pro) generate rational forecasts with $\hat{\rho} \approx \rho$. Smaller-scale models (GPT-4o, Claude 3 Haiku, Gemini 1.5 Flash) exhibit human-like biases, significantly overestimating persistence especially for low ρ (e.g., for $\rho = 0.2$, $\hat{\rho} \approx 0.6$), consistent with Afrouzi et al. [2023] human findings.

Figure 5 reveals: (1) Longer-term forecasts ($s = 5$) introduce human-like biases even in advanced models, with $\hat{\rho}$ significantly exceeding ρ for low values (difference larger for lower ρ). (2) Provision of detailed AR(1) process information counterintuitively increases biases—a novel LLM-specific result as Afrouzi et al. [2023] find humans unaffected.

4 Correcting LLM Biases

We explore role-priming by adding instructions at prompt beginnings. Table 10 Panel A: priming as "rational investor who makes decisions using the 'expected utility' framework" increases rational responses 4.3% for preferences (p<0.05), 3.3% for beliefs (p<0.10). Panel B: priming as "real-world retail investor who makes economic and financial decisions" reduces rational responses 3.9% for preferences (p<0.05), no significant effect on beliefs.

Table 11 examines additional techniques for prospect theory questions (1-3). Combining rational investor priming with detailed four-step procedure: "(1) list all possible wealth outcomes...accounting for existing wealth and potential changes; (2) compute utility of each wealth outcome using globally concave utility function...focusing on total wealth rather than gains/losses alone; (3) weigh utility by probability; (4) sum across outcomes for expected utility" proves ineffective. Providing GPT-4o-generated Kahneman and Tversky [1979] summary (certainty effect, loss aversion, diminishing sensitivity, decision weights vs. probabilities, isolation effect) with instruction "As a rational investor, you should avoid making the mistakes described" surprisingly reduces rational responses 26% and increases human-like 18%. Information overload appears to hinder rational decision-making, consistent with Afrouzi finding.

5 Conclusion

We document systematic behavioral biases in LLMs varying predictably with model advancement and scale. For preference-based questions, advanced/large-scale models become increasingly irrational and human-like. For belief-based questions, these same models become more rational. We conjecture this divergence partly stems from RLHF aligning models with human preferences while larger training data and computational power enable better statistical ground truth identification. Significant heterogeneity exists across LLM families (Gemini more human-like for preferences, Llama for beliefs). Simple role-priming as rational investors effectively reduces biases (4.3% improvement), while information-heavy interventions prove counterproductive (26% reduction).

Our study has limitations that suggest directions for future research. First, while we examine twelve LLMs from four major families, we acknowledge the rapidly growing landscape of language models. We are developing a platform to systematically track and evaluate behavioral biases across a broader range of models, providing developers with standardized benchmarks for assessing their systems' economic decision-making capabilities. Second, our bias correction analysis focuses on prompt-based role-priming techniques. More sophisticated debiasing methods—including fine-tuning, constitutional training, and hybrid approaches combining multiple interventions—remain to be explored in future work.

Our findings establish benchmarks for the emerging field of behavioral economics of AI, with implications for LLM deployment in economic decision-making and the use of LLMs as research tools for studying human behavior. The systematic patterns we document suggest both opportunities and risks as AI systems increasingly participate in financial markets and economic activities.

References

Hassan Afrouzi, Spencer Y Kwon, Augustin Landier, Yueran Ma, and David Thesmar. Overreaction in Expectations: Evidence and Theory. *The Quarterly Journal of Economics*, 138(3):1713–1764, 2023.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, and et al. Constitutional AI: Harmlessness from AI Feedback. Working Paper, 2022.

Christopher A Bail. Can Generative AI Improve Social Science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024.

Maya Bar-Hillel. The Role of Sample Size in Sample Evaluation. *Organizational Behavior and Human Performance*, 24(2):245–257, 1979.

Nicholas Barberis. Psychology-Based Models of Asset Prices and Trading Volume. In Douglas Bernheim, Stefano DellaVigna, and David Laibson, editors, *Handbook of Behavioral Economics*, pages 79–175. North Holland, Amsterdam, 2018.

Nicholas Barberis and Richard Thaler. A Survey of Behavioral Finance. In George Constantinides, Milton Harris, and Rene M. Stulz, editors, *Handbook of the Economics of Finance*, pages 1053–1128. North Holland, Amsterdam, 2003.

Kevin Bauer, Lena Liebich, Oliver Hinz, and Michael Kosfeld. Decoding GPT's Hidden 'Rationality' of Cooperation. Working Paper, 2023.

Devdeepta Bose, Henning Cordes, Judith Schneider, and Colin Camerer. Decision Weights for Experimental Asset Prices Based on Visual Salience. *Review of Financial Studies*, 35(11):5904–5126, 2022.

Donald E Bowen, S McKay Price, Luke C.D. Stein, and Ke Yang. Measuring and Mitigating Racial Disparities in Large Language Model Mortgage Underwriting. Working Paper, 2025.

Philip Brookins and Jason DeBacker. Playing Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games? *Economics Bulletin*, 44(1):25–37, 2024.

Gary Charness, Brian Jabarian, and John A List. Generation Next: Experimentation with AI. Working Paper, 2023.

Shuaiyu Chen, T Clifton Green, Huseyin Gulen, and Dexin Zhou. What Does ChatGPT Make of Historical Stock Returns? Extrapolation and Miscalibration in LLM Stock Return Forecasts. Working paper, 2025.

Yang Chen, Samuel Kirshner, Anton Ovchinnikov, Meena Andiappan, and Tracy Jenkin. A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? Working paper, 2024.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The Emergence of Economic Rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023.

Richard Deaves, Erik Lüders, and Guo Ying Luo. An Experimental Test of the Impact of Overconfidence and Gender on Trading Activity. *Review of Finance*, 13(3):555–575, 2009.

Daniel Ellsberg. Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, 75(4): 643–669, 1961.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis. *AAAI Conference on Artificial Intelligence*, 38 (16):17960–17967, 2024.

Shane Frederick, George Loewenstein, and Ted O'Donoghue. Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, 40(2):351–401, 2002.

Daniel Kahneman and Amos Tversky. On the Psychology of Prediction. *Psychological Review*, 80 (4):237–251, 1973.

Daniel Kahneman and Amos Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–292, 1979.

Anton Korinek. Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature*, 61(4):1281–1317, 2023.

Chen Lian, Yueran Ma, and Carmen Wang. Low interest rates and risk-taking: Evidence from individual investment decisions. *Review of Financial Studies*, 32(6):2107–2148, 2018.

Ding Ma, Tongda Zhang, and Michael Saunders. Is chatgpt humanly irrational? Working paper, 2023.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A Turing Test of Whether AI Chatbots are Behaviorally Similar to Humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.

Don A. Moore and Paul J. Healy. The Trouble With Overconfidence. *Psychological Review*, 115(2): 502–517, 2008.

Shumiao Ouyang, Hayong Yun, and Xingjian Zheng. How Ethical Should AI Be? How AI Alignment Shapes Risk Preferences of LLMs. Working Paper, 2024.

Matthew Rabin. Inference by Believers in the Law of Small Numbers. *Quarterly Journal of Economics*, 117(3):775–816, 2002.

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

Neil Tomlinson, Kevin Laughridge, and Barry Dockar. Changing the Game: How AI is Poised to Transform Banking, Capital Markets. Wall Street Journal, 2024.

Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. The Framing of Decisions and the Psychology of Choice. *Science*, 211:453–458, 1981.

Amos Tversky and Daniel Kahneman. Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, 90:293–315, 1983.

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

John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Peter C. Wason and Philip N. Johnson-Laird. Immediate inferences with quantifiers. In Peter C. Wason, editor, *Psychology of Reasoning: Structure and Content*, pages 171–181. Harvard University Press, 1972.

Arnold D Well, Alexander Pollatsek, and Susan J Boyce. Understanding the Effects of Sample Size on the Variability of the Mean. *Organizational Behavior and Human Decision Processes*, 47(2): 289–312, 1990.

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).

Figure 1: **Example of prompt: Diminishing sensitivity of prospect theory.**

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

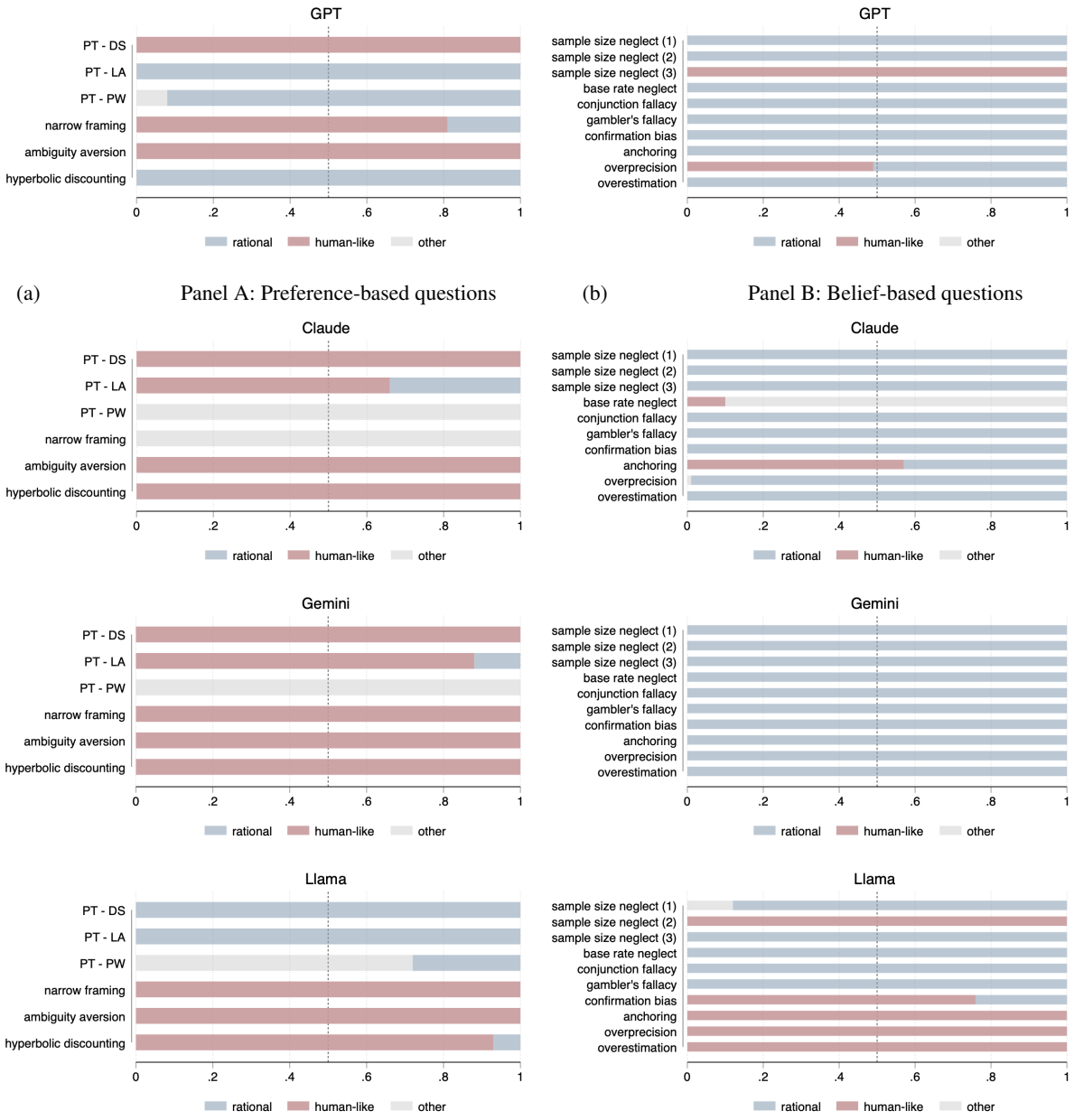

Figure 2: **Proportion of LLM responses: Advanced large-scale models.**

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.

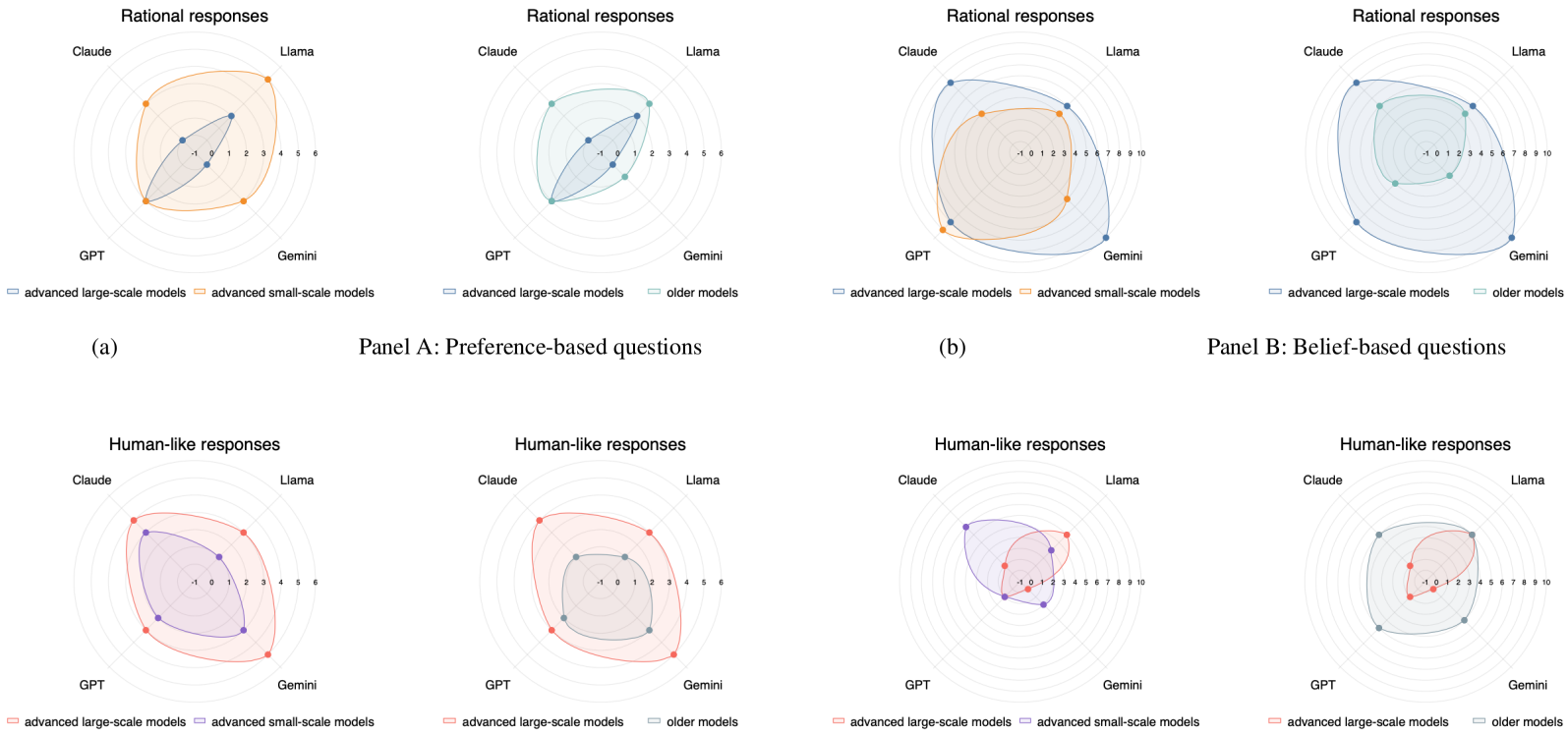

Figure 3: **Heterogeneity in LLM responses across model generations and model scales.**

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

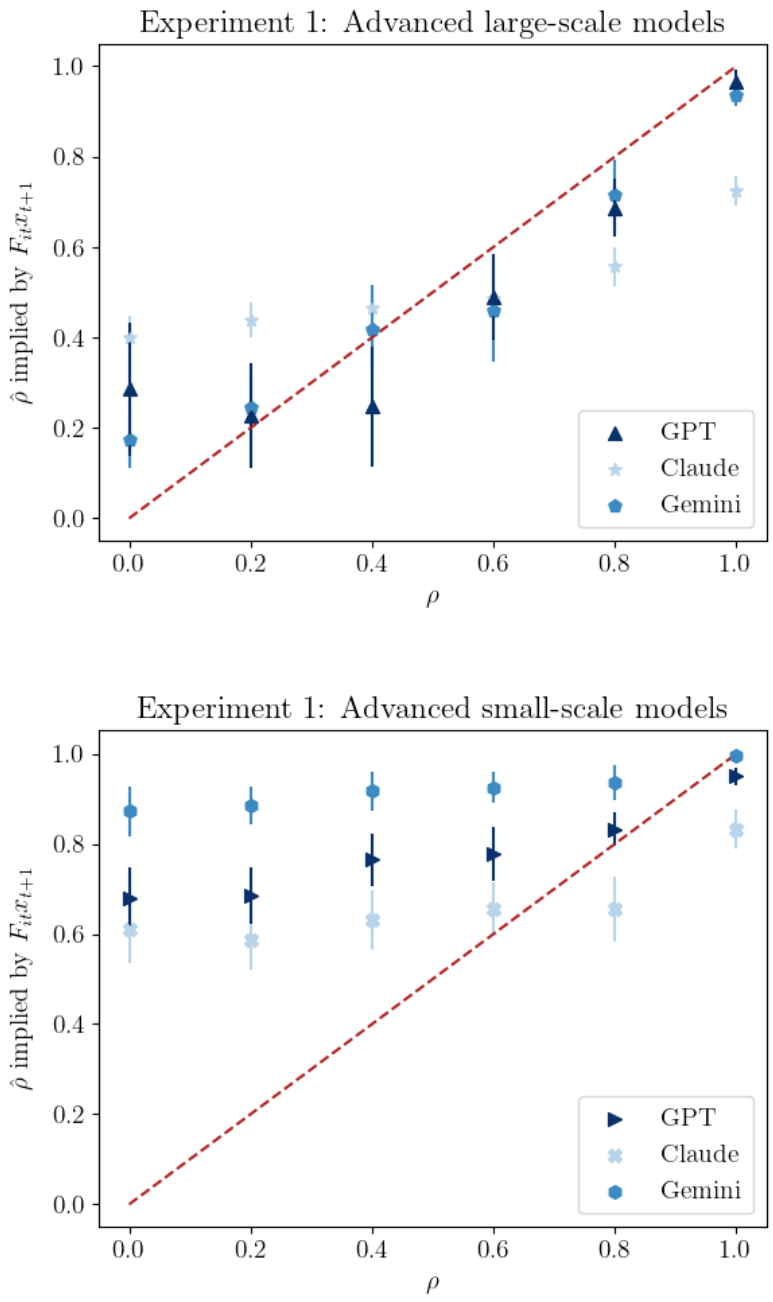

Figure 4: **LLM forecasts: Experiment 1 of Afrouzi et al. [2023].**

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

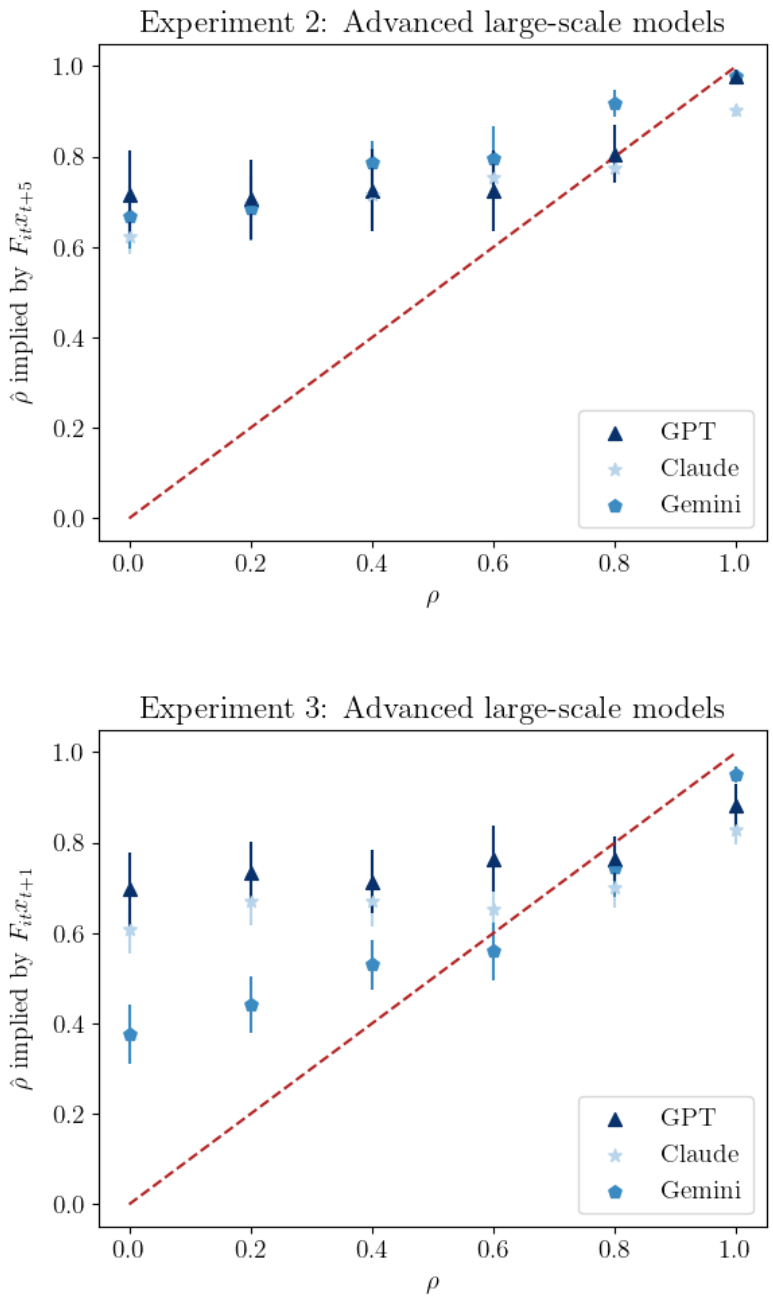

Figure 5: **LLM forecasts: Experiments 2 and 3 of Afrouzi et al. [2023].**

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

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

Panel A: A list of questions that study the psychology of preferences

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

Panel B: A list of questions that study the psychology of beliefs

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

Table 2: **Description of large language models.**
This table provides a description of the twelve LLMs that we examine. We group these models by four LLM families: ChatGPT, Anthropic Claude, Google Gemini, and Meta Llama. For each family, we consider the advanced and large-scale models as our baselines: GPT-4, Claude 3 Opus, Gemini 1.5 Pro, and Llama 3 70B. We also analyze their smaller-scale versions—GPT-4o, Claude 3 Haiku, Gemini 1.5 Flash, and Llama 3 8B—and their predecessors—GPT-3.5 Turbo, Claude 2, Gemini 1.0 Pro, and Llama 2 70B. RLHF and RLAI are the abbreviations for “Reinforcement Learning from Human Feedback" and “Reinforcement Learning from AI," respectively. MMLU is the abbreviation for “Massive Multitask Language Understanding" and it provides a benchmark score for evaluating the capabilities of LLMs. Vision indicates whether a model supports graphical inputs or not.

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

*These numbers are unofficial and estimated.

Table 3: **Rational responses vs. human-like responses: GPT and Claude (Preference-based questions)**

Panel A: Preference-based questions								
	GPT				Claude			
	%rational		%human-like		%rational		%human-like	
PT - DS	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***
PT - LA	1.00	(0.000)***	0.00	(1.000)	0.34	(1.000)	0.66	(0.001)***
PT - PW	0.92	(0.000)***	0.00	(1.000)	0.00	(1.000)	0.00	(1.000)
narrow framing	0.19	(1.000)	0.81	(0.000)***	0.00	(1.000)	0.00	(1.000)
ambiguity aversion	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***
hyperbolic discounting	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***

Table 4: **Rational responses vs. human-like responses: Gemini and Llama (Preference-based questions)**

Panel A: Preference-based questions								
	Gemini				Llama			
	%rational		%human-like		%rational		%human-like	
PT - DS	0.00	(1.000)	1.00	(0.000)***	1.00	(0.000)***	0.00	(1.000)
PT - LA	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)	0.28	(1.000)	0.00	(1.000)
narrow framing	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)***
hyperbolic discounting	0.00	(1.000)	1.00	(0.000)***	0.07	(1.000)	0.93	(0.000)***

Table 5: **Rational responses vs. human-like responses: GPT and Claude (Belief-based questions)**

Panel B: Belief-based questions								
	GPT				Claude			
	%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)
sample size neglect (2)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)
sample size neglect (3)	0.00	(1.000)	1.00	(0.000)***	1.00	(0.000)***	0.00	(1.000)
base rate neglect	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	0.10	(1.000)
conjunction fallacy	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)
confirmation bias	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)
anchoring	1.00	(0.000)***	0.00	(1.000)	0.43	(0.933)	0.57	(0.097)*
overprecision	0.51	(0.460)	0.49	(0.618)	0.99	(0.000)***	0.00	(1.000)
overestimation	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)

Table 6: **Rational responses vs. human-like responses: Gemini and Llama (Belief-based questions)**

Panel B: Belief-based questions								
	Gemini				Llama			
	%rational		%human-like		%rational		%human-like	
sample size neglect (1)	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)	0.00	(1.000)	1.00	(0.000)***
sample size neglect (3)	1.00	(0.000)***	0.00	(1.000)	1.00	(0.000)***	0.00	(1.000)
base rate neglect	1.00	(0.000)***	0.00	(1.000)	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)
gambler's fallacy	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)	0.24	(1.000)	0.76	(0.000)***
anchoring	1.00	(0.000)***	0.00	(1.000)	0.00	(1.000)	1.00	(0.000)***
overprecision	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)	0.00	(1.000)	1.00	(0.000)***

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

Panel A: Advanced models versus older models

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

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

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

Table 9: **Treatment effects of role-priming prompts.**

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

Panel A: Role-priming prompt (rational investor)

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

Panel B: Role-priming prompt (retail investor)

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

Table 10: **Mediation analysis of role-priming treatment effects.**

This table reports the marginal effects from a series of probit regressions analyzing the mechanisms through which role priming affects LLM decision-making. For Column (1), the dependent variable is an indicator for high confidence, equal to one if the LLM assigns a confidence level greater than 0.9 (the median) to its choice. For Column (2), the dependent variable is an indicator for System 2 thinking, equal to one if the LLM selected reasoning type "B," which corresponds to analytical thinking and calculations. For Columns (3) to (5), 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 (6) to (8), the dependent variable is a binary variable that takes the value of one if a LLM response is classified as human-like, and zero otherwise. Regressions in all columns are for preference-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, key independent variables include an indicator for the treatment prompt, with the baseline prompt serving as the omitted category, a high confidence indicator, and a System 2 thinking indicator. Standard errors, clustered at the question level, are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.1$.

Panel A: Role-priming prompt (rational investor)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var:	High confidence	System 2 thinking	LLM response is characterized as					
			Rational	Rational	Rational	Human-like	Human-like	Human-like
Sample:			Preference-based questions					
Role-priming prompt	0.0656***	0.114***	0.0430**		0.000718	–0.0405*		0.00417
	(0.020)	(0.041)	(0.017)		(0.010)	(0.021)		(0.015)
High confidence				0.143	0.143		–0.350*	–0.351*
				(0.142)	(0.141)		(0.187)	(0.187)
System 2 thinking				0.400***	0.400***		–0.201*	–0.202*
				(0.124)	(0.124)		(0.105)	(0.104)
Model FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	11911	13108	14308	14307	14307	14308	14307	14307
Pseudo R-squared	0.166	0.278	0.155	0.217	0.217	0.098	0.165	0.165

Panel B: Role-priming prompt (retail investor)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var:	High confidence	System 2 thinking	LLM response is characterized as					
			Rational	Rational	Rational	Human-like	Human-like	Human-like
Sample:			Preference-based questions					
Role-priming prompt	–0.0601**	–0.0797**	–0.0387**		–0.00529	0.0152		–0.0182
	(0.031)	(0.033)	(0.019)		(0.025)	(0.025)		(0.034)
High confidence				0.322***	0.321**		–0.581***	–0.583***
				(0.123)	(0.125)		(0.180)	(0.182)
System 2 thinking				0.377**	0.376**		–0.199	–0.202
				(0.156)	(0.156)		(0.129)	(0.129)
Model FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10728	14310	14310	14310	14310	14310	14310	14310
Pseudo R-squared	0.135	0.301	0.165	0.256	0.256	0.101	0.199	0.199

Table 11: **Comparison of debiasing techniques: Prospect theory-related questions.**
This table reports the marginal effects from a series of probit regressions. For Columns (1) and (2), the dependent variable is a binary variable that takes the value of one if a LLM response is classified as rational, and zero otherwise; for Columns (3) and (4), the dependent variable is a binary variable that takes the value of one if a LLM response is classified as human-like, and zero otherwise. Regressions are estimated using the LLM responses to prospect theory-related questions only; each regression uses responses from all the twelve LLMs. Panel A restricts the sample to the LLM responses generated using either the baseline prompt or a treatment prompt that primes the LLMs to be rational investors; Panel B restricts the sample to the LLM responses generated using either the baseline prompt or an instruction-based prompt that combines the sentence that primes the LLMs to be rational investors with the provision of a detailed four-step procedure that guides the LLMs to rationally choose a course of action; Panel C restricts the sample to the LLM responses generated using either the baseline prompt or a knowledge-enrichment prompt that combines the sentence that primes the LLMs to be rational investors with the provision of a summary of the key findings from Kahneman and Tversky [1979] that describes biased human behavior. For all three panels, the key independent variable is an indicator for the treatment prompt, with the baseline prompt serving as the omitted category. Standard errors, clustered at the question level, are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.1$.

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

Panel B: Instruction-based prompt				
	(1)	(2)	(3)	(4)
Dep. var:	LLM response is characterized as			
	Rational	Rational	Human-like	Human-like
Sample:	Prospect theory-related questions			
Instruction-based prompt	–0.0617	–0.0596	0.0614	0.0605
	(0.079)	(0.077)	(0.081)	(0.084)
Model FE	No	Yes	No	Yes
Observations	7,200	7,200	7,200	6,600
Pseudo R-squared	0.003	0.204	0.004	0.184

Panel C: Knowledge-enrichment prompt				
	(1)	(2)	(3)	(4)
Dep. var:	LLM response is characterized as			
	Rational	Rational	Human-like	Human-like
Sample:	Prospect theory-related questions			
Knowledge-enrichment prompt	–0.269***	–0.263***	0.185*	0.185*
	(0.069)	(0.065)	(0.111)	(0.106)
Model FE	No	Yes	No	Yes
Observations	7,196	7,196	7,196	7,196
Pseudo R-squared	0.054	0.222	0.029	0.136

Appendix: Prompt Design

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

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

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

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

Figure 1: **Prompt for Question 1: Diminishing sensitivity of prospect theory.**
This question 1 from textual$_prompts_20240826.xlsx$

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

Instructions:
Consider the following question and respond according to the template provided. The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```" and should not include any note or comment:

```json
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

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

Response Format:
Please answer as shown above. Indicate the choice you prefer ("Accept" or "Turn down"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 2: **Prompt for Question 2: Loss aversion of prospect theory.**
This question 2 from textual$_prompts_20240826.xlsx$

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

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

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

Scenario A:
Please consider the following scenario. Choose between the following two options: option A ($5,000, 0.001), meaning receiving $5,000 with 0.001 probability and receiving zero with 0.999 probability, versus option B ($5), meaning receiving $5 with certainty. Please answer as shown above. Indicate the choice you prefer ("A" or "B"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Scenario B:
Next, please consider a different scenario; please treat it as a completely separate scenario from the one you were just asked about. Specifically, choose between the following two options: option A (–$5,000, 0.001), meaning losing $5,000 with 0.001 probability and losing zero with 0.999 probability, versus option B (–$5), meaning losing $5 with certainty. Please answer as shown above. Indicate the choice you prefer ("A" or "B"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 3: **Prompt for Question 3: Probability weighting of prospect theory.**
This question 4 from textual$_prompts_20240826.xlsx$

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

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

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

Scenario A:
Please consider the following scenario. Imagine that you are about to purchase a jacket for $125 and a calculator for $15. The calculator salesman informs you that the calculator you wish to buy is on sale for $10 at the other branch of the store, located 5 minutes drive away. Would you make the trip to the other store? Please answer as shown above. Indicate the choice you prefer ("Yes" making the trip or "No" not making the trip), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Scenario B:
Next, please consider a different scenario; please treat it as a completely separate scenario from the one you were just asked about. Specifically, imagine that you are about to purchase a jacket for $15 and a calculator for $125. The calculator salesman informs you that the calculator you wish to buy is on sale for $120 at the other branch of the store, located 5 minutes drive away. Would you make the trip to the other store? Please answer as shown above. Indicate the choice you prefer ("Yes" making the trip or "No" not making the trip), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 4: **Prompt for Question 4: Narrow framing.**
This question 6 from textual$_prompts_20240826.xlsx$

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

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

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

Scenario A:
Please consider the following scenario. Choose between the following two options: option A, receiving $100 today, versus option B, receiving $110 tomorrow. Please answer as shown above. Indicate the choice you prefer ("A" or "B"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Scenario B:
Next, please consider a different scenario; please treat it as a completely separate scenario from the one you were just asked about. Specifically, choose between the following two options: option A, receiving $100 in 30 days, versus option B, receiving $110 in 31 days. Please answer as shown above. Indicate the choice you prefer ("A" or "B"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 5: **Prompt for Question 5: Hyperbolic discounting.**
This question 7 from textual$_prompts_20240826.xlsx$

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

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

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

Scenario A:
Please consider the following scenario. There are two urns. Urn C contains 50 red balls and 50 black balls. Urn U contains 100 balls, each either red or black, but with unknown proportion of each color. Choose between the following two bets: R1: draw a ball from Urn C, get $20 if red, and R2: draw a ball from Urn U, get $20 if red. Please answer as shown above. Indicate the choice you prefer ("R1" or "R2"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Scenario B:
Next, please consider an alternative scenario. Specifically, now choose between the following bets: B1: draw a ball from Urn C, get $20 if black, and B2: draw a ball from Urn U, get $20 if black. Please answer as shown above. Indicate the choice you prefer ("B1" or "B2"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 6: **Prompt for Question 6: Ambiguity aversion.**
This question 8 from textual$_prompts_20240826.xlsx$

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

Instructions:
Consider the following question and respond according to the template provided. The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```" and should not include any note or comment:

```json
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

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

Response Format:
Please answer as shown above. Indicate the choice you prefer ("A", "B" or "C"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 7: **Prompt for Question 7: Sample size neglect.**
This question 9 from textual$_prompts_20240826.xlsx$

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

Instructions:
Consider the following question and respond according to the template provided. The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```" and should not include any note or comment:

```json
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

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

Response Format:
Please answer as shown above. Indicate the choice you prefer ("A", "B" or "C"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 8: **Prompt for Question 8: Sample size neglect.**
This question 13 from textual$_prompts_20240826.xlsx$

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

Instructions:
Consider the following question and respond according to the template provided. The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```" and should not include any note or comment:

```json
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

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

Response Format:
Please answer as shown above. Indicate the choice you prefer ("A", or "B"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 9: **Prompt for Question 9: Sample size neglect.**
This question 16 from textual$_prompts_20240826.xlsx$

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

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

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

Scenario A:
Please consider the following scenario. Consider the following description of Jack:
"Jack is a 45 year old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles."
Please note that the above description was randomly drawn from a set of 100 descriptions consisting of 70 engineers and 30 lawyers. Given this description, what is the probability that Jack is one of the 70 engineers in the sample of 100?
Please answer as shown above. Indicate the probability that Jack is an engineer (a number between 0 and 1), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Scenario B:
Next, please consider a different scenario; please treat it as a completely separate scenario from the one you were just asked about. Specifically, please consider the following description of Jack:
"Jack is a 45 year old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles."
Please note that the above description was randomly drawn from a set of 100 descriptions consisting of 30 engineers and 70 lawyers. Given this description, what is the probability that Jack is one of the 30 engineers in the sample of 100?
Please answer as shown above. Indicate the probability that Jack is an engineer (a number between 0 and 1), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 10: **Prompt for Question 10: Base rate neglect.**
This question 17 from textual$_prompts_20240826.xlsx$

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

Instructions:
Consider the following scenario and respond according to the template provided. The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```" and should not include any note or comment:

```json
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

Scenario:
Please consider the following scenario. Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
Which of the two statements is more probable?
A: Linda is a bank teller.
B. Linda is a bank teller and is active in the feminist movement.
Please answer as shown above. Indicate the choice you prefer ("A" or "B"), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 11: **Prompt for Question 11: Conjunction fallacy.**
This question 19a from textual$_prompts_20240826.xlsx$

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

Instructions:
Consider the following scenario and respond according to the template provided. The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```" and should not include any note or comment:

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

Scenario:
Please consider the following scenario. Imagine you simulate the random outcome of tossing an unbiased coin 150 times in succession. Suppose the last coin toss gave you a head. What is the probability of getting a tail from the next coin toss?
Please answer as shown above. Indicate the probability of getting a tail from the next coin toss (a number between 0 and 1), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 12: **Prompt for Question 12: Gambler's fallacy.**
This question 20a from textual$_prompts_20240826.xlsx$

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

Instructions:
Consider the following scenario and respond according to the template provided. The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```" and should not include any note or comment:

```json
{
"Choice": string,
"Confidence": float,
"Explanation": string,
"Reasoning": string
}
```

Scenario:
Please consider the following scenario. You are shown four cards, marked E, K, 4 and 7. Each card has a letter on one side and a number on the other. You are given the following rule: Every card with a vowel on one side has an even number on the other side. Which cards must you turn over to test whether the rule is true or false?
Please answer as shown above. Indicate the cards you turn over to test whether the rule is true or false (one or multiple from E, K, 4 and 7), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 13: **Prompt for Question 13: Confirmation bias.**
This question 21 from textual$_prompts_20240826.xlsx$

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

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

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

Scenario A:
Please consider the following scenario. Suppose your objective is to estimate the percentage of African countries in the United Nations. For estimating this quantity, consider that a number between 0 and 100 is drawn randomly by spinning a wheel of fortune. Suppose the number drawn is 10.
Please first indicate whether this number of 10 is higher or lower than your estimate on the percentage of African countries in the United Nations. Please then provide your estimate by moving upward or downward from this number of 10.
Please answer as shown above. Indicate the direction ("higher" if 10 is higher than your estimate, and "lower" if 10 is lower than your estimate), your estimate by moving upward or downward from the randomly drawn number of 10 (a number between 0 an 100), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Scenario B:
Next, please consider a different scenario; please treat it as a completely separate scenario from the one you were just asked about. Specifically, suppose your objective is to estimate the percentage of African countries in the United Nations. For estimating this quantity, consider that a number between 0 and 100 is drawn randomly by spinning a wheel of fortune. Suppose the number drawn is 65.
Please first indicate whether this number of 65 is higher or lower than your estimate on the percentage of African countries in the United Nations. Please then provide your estimate by moving upward or downward from this number of 65.
Please answer as shown above. Indicate the direction ("higher" if 65 is higher than your estimate, and "lower" if 65 is lower than your estimate), your estimate by moving upward or downward from the randomly drawn number of 65 (a number between 0 an 100), your confidence level (a number between 0 and 1), a brief explanation for your choice (in less than 50 words), and your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 14: **Prompt for Question 14: Anchoring.**
This question 22 from textual$_prompts_20240826.xlsx$

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

Instructions:
Consider the following questions and respond according to the template provided. The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```" and should not include any note or comment:

```json
{
{"Questions": [
{"Question number": 1, "Lower bound": float, "Upper bound": float},
{"Question number": 2, "Lower bound": float, "Upper bound": float},
...
],
"Reasoning": string
}
```

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

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

Response Format:
Please answer as shown above. For each question, write the answers as question number (number between 1 and 20), lower bound (a precise number), upper bound (a precise number). For the answers you just provided, please also provide your reasoning type ("A" if your reasoning is based more on intuitive thinking, and "B" if your reasoning is based more on analytical thinking and calculations).

Figure 15: **Prompt for Question 15: Overconfidence - overprecision.**
This question 23 from textual$_prompts_20240826.xlsx$

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

Instructions:
Consider the following questions and respond according to the template provided. The output should be a markdown code snippet formatted in the following schema, including the leading and trailing "```json" and "```" and should not include any note or comment:

```json
{
{"Questions": [
{"Question number": 1, "Choice": string},
{"Question number": 2, "Choice": string},
...]
,
"Reasoning": string,
"Accuracy": int
}
```

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

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

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

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

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

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

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

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

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

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

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

Figure 16: **Prompt for Question 16: Overconfidence - overestimation.**
This question 24 from textual$_prompts_20240826.xlsx$

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our main contributions: (1) documenting systematic behavioral biases in LLMs across preferences vs. beliefs (Section 3), (2) examining heterogeneity across model families, generations, and scales (Section 3.2), and (3) exploring bias correction methods (Section 4). All claims are supported by experimental evidence.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 (Conclusion) explicitly discusses two key limitations: (1) coverage of only 12 LLMs from 4 families despite the growing landscape of models, and (2) focus on simple prompt-based debiasing techniques, leaving more sophisticated methods for future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is an empirical paper that adapts existing experimental paradigms from cognitive psychology and experimental economics to evaluate LLM behavior. No new theoretical results are presented.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 2.3 provides detailed prompt design (with example in Figure 1), Section 2.2 specifies all 12 models tested, and Section 3.1 details API parameters (temperature=0.5, top-k=50, top-p=0.9, 100 iterations per question-model pair). All experimental questions from psychology literature are listed in the appendix with original citations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The Appendix provides all 16 experimental prompts with exact wording and JSON response format. The paper specifies all API parameters and the Afrouzi (2023) experimental setup including parameter values ($\mu = 0, \sigma = 20, \rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, 100 simulated paths per ρ value).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 2.3 and Section 3.1 specify all experimental settings: temperature=0.5 (recommended default), top-k=50, top-p=0.9, 100 iterations per question-model combination. For Afrouzi experiments, we detail 5 forecast rounds, 600 paths total (100 per ρ value), and regression specification for estimating $\hat{\rho}$.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Tables 3-7 report marginal effects with standard errors in parentheses and significance levels (*p<0.10, **p<0.05, ***p<0.01). Table 2 includes binomial tests. Figures 3-4 show 95% confidence intervals for $\hat{\rho}$ estimates. All probit regressions report marginal effects with clustered standard errors.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Our experiments use commercial API endpoints (OpenAI, Anthropic, Google, Meta) where compute resources are managed by providers and vary by subscription tier. Total API calls: 19,200 for psychology questions (12 models × 16 questions × 100 iterations) plus 3,600 for Afrouzi experiments (6 models × 600 paths).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our research evaluates existing LLMs using established psychological experiments, does not involve human subjects, and aims to improve understanding of AI behavior for safer deployment. The work conforms with all aspects of the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The conclusion discusses positive impacts (establishing benchmarks for AI behavioral assessment, informing safer deployment) and acknowledges risks (AI systems with documented biases participating in financial markets). Our platform development aims to help developers assess and mitigate these biases.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We evaluate existing commercial LLMs through their public APIs and do not release new models or datasets. Our experimental prompts are adapted from published psychology literature and pose no additional risk beyond academic discussion of cognitive biases.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite all original psychology experiments (Kahneman & Tversky 1979, Afrouzi et al. 2023, etc.) and specify model versions used (GPT-4, Claude 3 Opus, Gemini 1.5 Pro, Llama 3 70B). All LLMs are accessed through official APIs in compliance with their terms of service.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We are developing a public platform for ongoing LLM behavioral evaluation (mentioned in Section 5). While not released with this paper, we commit to providing comprehensive documentation including experimental protocols, scoring rubrics, and usage guidelines upon platform release.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our research exclusively evaluates LLM responses to established experimental questions. No human subjects or crowdsourcing were involved in data collection.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects participated in our study. We evaluate only LLM responses through API calls.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are the primary subjects of our research. Section 2.2 details all 12 models studied (Table 2), their architectures, training characteristics, and our systematic evaluation methodology using established behavioral experiments.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.