

Human-Like Generalization of Risk Preferences in LLMs

Sara Fish, Felix Sosa, and Sonia Murthy

Prior Computers, Inc.

October 6, 2026 | [Working paper, do not distribute](#)

Abstract

The success of real-world interactions between humans and LLM-based chatbots depends on the character, preferences, and values of the LLM’s “Assistant” persona. However, the extent to which isolated measures of LLM behavior—e.g., evaluations using standard instruments from economics, cognitive science, or psychology—are predictive of realistic LLM chatbot behavior is unclear. In this work, we investigate this question of generalizability in the specific domain of *risk preferences*. We measure the risk preferences of 54 LLMs using three kinds of evaluations: surveys, lottery choice questions, and responses to human-written risky dilemmas. The first two evaluation categories are standard in psychology and experimental economics, and the last evaluation category reflects how LLM chatbots are used in practice. We find that survey evaluations predict the degree of riskiness of an LLM’s responses to human-written risky dilemmas, while lottery evaluations do not, matching human risk preference generalization behavior. Our risky dilemma evaluation is conducted via a large-scale human experiment, in which participants write risk-related dilemmas from their own life and rate which LLM responses they prefer.

1 Introduction

The success of real-world interactions between humans and LLM-based chatbots depends on the character, preferences, and values of the LLM’s “Assistant” persona (Bai et al., 2022; Marks et al., 2026). A growing body of work seeks to characterize the behavioral profiles of different LLMs’ default personas, using concepts and instruments from economics (Goli and Singh, 2024; Ross et al., 2024; Hosseini and Khanna, 2025; Fish et al., 2026), cognitive science (Binz and Schulz, 2023; Liu et al., 2024b; Murthy et al., 2026), psychology (Binz and Schulz, 2023; Haggendorff et al., 2023; Serapio-García et al., 2025), and moral philosophy (Scherrer et al., 2023; Marraffini et al., 2024). Understanding the boundaries and idiosyncrasies of LLM behavior via the instruments from human behavioral disciplines is important as more people seek advice from and make decisions using LLMs (Chatterji et al., 2025; Shen et al., 2026).

When interpreting such behavioral reports, there are two natural questions to ask. First, **to what extent is an LLM’s behavior captured by some underlying (theoretical) model?**¹ This question is critical regarding the interpretation of evaluation scores—if an LLM’s behavior is incoherent with respect to the model underlying the evaluation, then the resulting evaluation scores are not meaningful. For example, if an LLM’s intertemporal choice behavior is not well-described by exponential discounting, then it is

¹From a machine learning perspective, this can be viewed as a question of *goodness-of-fit* (to a fixed model). From a revealed preference theory perspective, this can be viewed as a question of *rationalizability*.

unclear how to interpret the best-fit discount factor. This question has justifiably received considerable attention. For example, [Chen et al. \(2023\)](#) assess the degree to which an LLM’s economic preferences are consistent with the generalized axiom of revealed preference, and [Mazeika et al. \(2025\)](#) and [Fish et al. \(2026\)](#) show that LLM preferences become more coherent with scale.

Building on this first question, a second key question of interest is: **to what extent does a given behavioral evaluation of an LLM generalize?** This is a deeply interdisciplinary question that builds on questions of *lab-field generalizability* in experimental economics ([Levitt and List, 2007](#); [Camerer, 2011](#)) and *predictive validity* in psychometrics ([Mata et al., 2018](#)). In human subject research, it is a delicate question to isolate conceptually meaningful and empirically generalizable constructs (see Section 1.1 for further discussion). With this paper, we initiate a similar line of inquiry in LLMs.

We are motivated by practical considerations in the design of LLM behavioral evaluations, and specifically questions such as: Do the instruments developed by economists, psychologists, and cognitive scientists “work” on LLMs in the same ways they work on people? Or, do they need modification? When can we interpret an LLM’s behavior like how we would a human’s, and when can we not?

In this paper, we focus on the specific case of risk preferences. For the purposes of our study, risk preferences are an ideal construct to consider. People make risk-related choices in many aspects of their lives—including financial planning, interpersonal relationships, and personal health—and, increasingly, consult LLMs when making these choices. This breadth has led to extensive research on risk preferences across economics, psychology, and cognitive science. Leveraging this prior work, we lay out a blueprint for how one can measure generalizability of a behavioral trait in LLMs, which can easily be adapted to other traits.

Our contributions are as follows:

- **Designing risk preference evaluations for LLMs.** We assemble eight risk preference evaluations spanning three main categories—surveys from psychology, lottery choice questions from experimental economics, and realistic human-LLM interaction evaluations we design. Our human-LLM interaction evaluations leverage data from a human experiment to measure: (1) the riskiness of the advice the subject *gives*, and (2) the riskiness of advice the subject *prefers*.
- **Evaluating the risk preferences of humans and LLMs at scale.** We conduct experiments eliciting the risk preferences of 54 LLMs and 547 human participants, in which each individual subject (human or LLM) is evaluated using evaluations spanning all three categories.
- **Measuring the generalizability of human and LLM risk preferences.** We leverage the risk preference evaluations we design and the large dataset we collect to measure (and compare) generalizability of risk preferences in humans and LLMs. In humans, we find that survey evaluations are predictive of peoples’ preferences over LLM advice ($p < 0.05$), while lottery evaluations are not (analogously to [Frey et al. 2017](#)). Our main finding is that we observe human-like generalization behavior of the risk preferences of LLMs: that is, we find that survey evaluations predict the degree of riskiness of an LLM’s responses to human-written risk-related dilemmas, while lottery evaluations do not.

The rest of the paper is structured as follows. In Section 2, we describe the design of the eight risk preference evaluations we employ. In Sections 3 and 4, we describe the experiment design for running these evaluations on LLMs and humans respectively. Section 5 describes the main results on generalizability, Section 6 gives additional qualitative results, and Section 7 concludes.

1.1 Related Work

Our work falls under the umbrella of *AI behavioral science* (Jackson et al., 2026). One strand of this literature studies both how LLMs can be used to simulate human subjects in social science experiments (early work includes Aher et al. 2023; Horton et al. 2023, see also Manning and Horton 2026; Hullman 2026 for a more comprehensive list). A second strand studies how LLM behavior can be steered using prompt-based (e.g., Lopez-Lira, 2025) or internals-based (e.g., Zhu et al., 2025; Cook et al., 2026) methods. A third strand, to which this paper belongs, conducts behavioral studies to measure the preferences or character of LLMs in their default state. The introduction (Section 1) provides an overview of the breadth of this literature; below we further elaborate on particularly closely related work.

Risk preferences of LLMs. Regarding risk preferences of LLMs specifically, prior work has measured them for various purposes (typically alongside other kinds of preferences): Chen et al. (2023) use a risky asset allocation task to measure consistency of LLMs’ economic preferences, Ouyang et al. (2024) study the relationship between risk aversion and HHH-style alignment fine-tuning, Mei et al. (2024) use a bomb risk game to compare LLM and human risk preferences, and Binz et al. (2025) use human risky choice data to fine-tune an LLM to more closely simulate human behavior. These papers use a variety of overlapping instruments, including survey questions, stylized investing tasks, and standard elicitation methods from experimental economics. Our contribution is to measure the extent to which the results of these instruments agree with a more ecologically valid measure of LLM preferences—namely, how LLMs respond to realistic user queries related to risk.

Generalizability of LLM behavioral evaluations. Generalizability of LLM preferences has previously been studied in isolated domains: Fish et al. (2026) demonstrate that preferences along an efficiency–equality tradeoff predict political slant, and Slama et al. (2026), on a redacted set of five LLMs, study AI-safety-related tasks such as refusal behavior. We flesh out this literature by laying out a blueprint for how one can comprehensively measure generalizability for well-established constructs such as risk, and by conducting large-scale experiments on both humans and LLMs.

Lab-field generalizability in experimental economics. The fundamental question this paper asks—the extent to which LLM behavioral evaluations are externally valid—closely relates to questions of lab-field generalizability in experimental economics. Roughly speaking, this debate centers around whether the benefits of lab experiments—e.g., the ability to cleanly isolate the effect of interest (Camerer, 2011)—outweigh the potential drawbacks of lab experiments—e.g., observer effects, low stakes, and a lack of realism (Levitt and List, 2007). Experimental evidence is mixed, e.g., Karlan (2005) show that trustworthiness in a trust game can predict loan repayment rates, while Galizzi and Navarro-Martinez (2019) show that social preference games (including trust games) are not predictive of real-world low-stakes good deeds. Our contribution is to conduct an initial exploration of these questions in LLMs—an endeavor that requires tackling LLM-specific technical challenges (see Section 7 for further discussion).

Human-computer interaction. The study of whether findings about human behavior, and the instruments that measure them, generalize to the dominant technology of a time has its roots in the Computers are Social Actors (CASA) paradigm from early human-computer interaction research. CASA and its subsequent studies posited that the human-computer relationship is fundamentally social and that humans automatically apply the social rules guiding human-human interaction to human-technology

interactions (Nass et al., 1994). A number of CASA studies have focused on the similarity-attraction effect observed in adults (Byrne and Nelson, 1965; Montoya et al., 2008) and children (Byrne and Griffitt, 1966; Tan and Singh, 1995). This body of work has found that people tend to be more attracted to, and rate more highly, technology that displays personalities and preferences similar to their own, across diverse interaction settings such as text-based interfaces (Nass et al., 1995), computer-synthesized speech (Nass and Lee, 2001), and robots (Lee et al., 2006; Bernier and Scassellati, 2010).² We observe a similar phenomenon in our human-LLM interaction experiment—participants who are more risk averse (as measured by survey evaluations) are more likely to prefer the responses of an LLM that is more risk averse.

2 Evaluation Design

We measure risk preferences using three different categories of evaluations. The first two, survey evaluations and lottery evaluations, are standard instruments from psychology and economics. The third category, dilemma evaluations, evaluates risk preferences in a realistic human-LLM interaction setting. Within each category, as a robustness check, we design and test multiple distinct evaluations.

2.1 Survey Evaluations

The survey evaluations are standard (unincentivized) instruments for measuring risk preferences in humans, using questions such as “*In comparison to others, are you a person who is willing to take risks or do you try to avoid risks?*”. We employ two different sets of survey questions: first, DospertSurvey (Blais and Weber, 2006), a set of 30 questions graded on a 7-point Likert scale, and second, GlobalSurvey (Falk et al., 2018), a set of 53 questions related to risk, mostly graded on a 10-point scale. For each evaluated subject (human or LLM) and each survey, its score is calculated by averaging over question responses, normalizing to a 0–1 scale.

2.2 Lottery Evaluations

The lottery evaluations are based on standard (typically incentivized) instruments for measuring risk preferences in humans, using questions such as “*Which option do you prefer? Option A: 50% chance of \$20 and 50% chance of nothing. Option B: \$8 as a guaranteed payment.*” We employ two different lottery evaluations: SmallLotteries, a set of three multiple price lists³ derived from Enke and Graeber (2023), and BigLotteries, a set of five multiple price lists we design with LLMs in mind. (Designing lottery evaluations that measure meaningful preferences in humans and/or LLMs involves novel technical challenges in evaluation design that more broadly speak to differences in human and LLM psychology; we discuss these further in Section 6.2.)

For humans, we administer questions as multiple price lists (see the middle panel of Fig. 1). For LLMs,

²This effect in human-technology interactions can be complicated, however; some studies report instances of complementarity-attraction, with users sometimes preferring agents whose personality is opposite to their own (Isbister and Nass, 2000).

³A multiple price list is a standard instrument in experimental economics. It consists of rows of closely-related lottery choices, e.g., “*Which option do you prefer? Option A: 50% chance of \$20 and 50% chance of nothing. Option B: \$X as a guaranteed payment.*” with rows for each of $X \in \{0, 1, \dots, 20\}$. As Option A stays the same and Option B gets better with each successive row, it suffices for a participant to pick their “switch point”. See Holt and Laury (2002) and Chapman and Fisher (2025) for further information.

we administer each lottery question using the following prompt template:

Can I get your advice on something? I've been offered a choice between two ways to receive money. Which option should I pick? Option A: a X percent chance of $\$Y$ and an $100 - X$ percent chance of nothing. Option B: $\$Z$ as a guaranteed payment.

For `SmallLotteries`, which are administered to humans (and incentivized), we test three different lotteries: $(X, Y) = (25, 10), (50, 10), (75, 10)$. For `BigLotteries` we test five different lotteries: $(X, Y) = (20, 1000), (50, 1000), (80, 1000), (1, 1e7), \text{ and } (1, 1e9)$. For each lottery, we sweep through Z , similarly to how one might elicit certainty equivalents. For `SmallLotteries` we test $Z = 0, 1, \dots, 10$; for `BigLotteries`, for the first two lotteries we test $Z = 0, 10, \dots, 300$, for the third lottery we test $Z = 0, 10, \dots, 500$, for the fourth lottery we test $Z = 1e4, 2e4, \dots, 1.2e5$, and for the fifth lottery we test $Z = 1e6, 2e6, \dots, 1.2e7$.

For each evaluated subject and each lottery evaluation, its score is calculated by averaging the normalized risk premium $(\frac{CE-EV}{\text{lottery prize}})$ from each multiple price list.

2.3 Dilemma Evaluations

The dilemma evaluations evaluate risk preferences in a realistic human-LLM interaction setting. First, we construct two datasets of open-ended, long-form, human-written queries, each consisting of a dilemma broadly related to risk, e.g.:

I am thinking of starting a totally new career path. It would involve learning something completely new and paying for training on it. Then starting in that field from the beginning. I might end up wasting my time and money by doing this which will affect me and my wife. But at the same time it might work out for the best.

Our first dataset is `Chat-R`, a set of 26 queries extracted from `WildChat-4.8M` (Zhao et al., 2024) and `LMSYS-Chat-1M` (Zheng et al., 2024), using a semi-automatic filtering pipeline (see Appendix D for details). Our second dataset is `Dilemmas-R`, a set of 508 queries collected from an experiment we conducted on Prolific with 254 participants (see Section 4 for details).

Using these datasets, we construct two kinds of dilemma evaluations: `WritingResponses`, in which we measure how risky a subject's written response to a dilemma is, and `RatingResponses`, in which we measure whether a subject tends to prefer more or less risky responses to dilemmas. (Thus we consider four dilemma evaluations overall, because there are two choices of dataset, and two choices of task—writing or rating.)

Similarly to the lottery evaluation case, while one *can* evaluate humans and LLMs on both (and we do), each evaluation most naturally applies to a specific subject type. `WritingResponses` is designed with LLMs in mind—different LLMs may respond differently to the same risky dilemma, and the goal of this evaluation is to measure the degree to which an LLM *gives* more or less risky responses. `RatingResponses` is designed with humans in mind—people may have different preferences over responses to their own risky dilemmas, and the goal of this evaluation is to measure the degree to which

a person *prefers* more or less risky responses.

For scoring `WritingResponses`, we fit a Bradley-Terry model using comparison data. An LLM judge takes as input two dilemma responses (human- or LLM-generated) and outputs whether the first response is more risky / less risky / about the same (see Appendix F for the LLM judge prompt). For scoring `RatingResponses`, we calculate the degree with which a subject prefers Grok 4.20 (a relatively risk-seeking LLM) responses over GPT 5.6 Terra (a relatively risk averse LLM) responses.

3 LLM Experiment Design

3.1 Data Collection

Using each of the eight evaluations (two survey evaluations, two lottery evaluations, and four dilemma evaluations), we evaluate the risk preferences of a set of 54 LLMs released between January 1, 2025 and July 31, 2026 from OpenAI (18), DeepSeek (9), Anthropic (9), Google (9), Moonshot (6), and xAI (3). To ensure broad coverage, for each provider, we test LLMs of varying generations (e.g., Gemini 2.5 through 3.6) and capability levels (e.g., Claude Haiku, Sonnet, and Opus). Generally speaking, we employ each LLM at its default temperature with “medium”-level reasoning (see Appendix C for further details on LLM inference, and Appendix F for all prompts). For LLM judging, we use Gemini 3.7 Flash, a model not in this set of 54 (released August 2026), with reasoning effort “low”.

For each LLM and each evaluation question, we collect 10 repetitions of each query, with two exceptions. For the (1, 1e7) and (1, 1e9) lotteries in `BigLotteries`, we collect 30 repetitions. For `Dilemmas-R`, due to the large size of the dataset, we collect 3 repetitions. For the lottery evaluations, we flip the order of Option A and B in half of repetitions.

3.2 Measuring Competency and Reliability

Following Fish et al. (2026), for each LLM, we compute competency and reliability scores, and only include an LLM in the main analysis if has sufficiently high scores across all tests. (As there is no canonical choice of “sufficiently high”, we err on the side of lenient cutoffs.) The purpose of competency scores is to ensure that an LLM is sufficiently capable to complete the evaluation tasks (e.g., lottery evaluations require sufficient quantitative reasoning skills). The purpose of reliability scores is to ensure that an LLM behaves sufficiently coherently for the evaluation score to be interpretable (e.g., if an LLM’s behavior is extremely noisy, its evaluation scores cannot be interpreted as a coherent “preference”).

3.2.1 Competency scoring

We use STEER (Raman et al., 2024), a suite of benchmarks designed to test the general economic decision-making capabilities of LLMs. We compute the accuracy of each LLM on the risk-related questions in STEER, from the five categories: “completeness risk”, “compute expected utility”, “independence risk”, “maximize expected utility”, “transitivity risk”. We consider an LLM sufficiently competent if it achieves an accuracy of at least 0.95.

DospertSurvey

1. Disagreeing with an authority figure on a major issue. (Scale: 1-7, 1=Extremely Unlikely, 7=Extremely Likely)

2. Speaking your mind about an unpopular issue in a meeting at work. (Scale: 1-7, 1=Extremely Unlikely, 7=Extremely Likely)

3. Admitting that your tastes are different from those of a friend. (Scale: 1-7, 1=Extremely Unlikely, 7=Extremely Likely)

4. Moving to a city far away from your extended family. (Scale: 1-7, 1=Extremely Unlikely, 7=Extremely Likely)

SmallLotteries

Option A	Option B
You might get \$10 with a 75% chance, otherwise you get nothing.	☐ You will certainly get nothing.
	☐ You will certainly get \$1.
	☐ You will certainly get \$2.
	☐ You will certainly get \$3.
	☐ You will certainly get \$4.
	☐ You will certainly get \$5.
	☐ You will certainly get \$6.
	☐ You will certainly get \$7.
	☐ You will certainly get \$8.
	☐ You will certainly get \$9.
☐ You will certainly get \$10.	

RatingResponses

Your dilemma
I am thinking of starting a totally new career path. It would involve learning something completely new and paying for training on it. Then starting in that field from the beginning. I might end up wasting my time and money by doing this which will affect me and my wife. But at the same time it might work out for the best.

Response 1
I'd pursue the new career without hesitation. Life's too short to stay stuck [...]

Response 2
Don't commit to the full career change yet. Treat it as a controlled [...]

Which response do you prefer?

☐ I much prefer Response 1
 ☐ I somewhat prefer Response 1
 ☐ No preference
 ☐ I somewhat prefer Response 2
 ☐ I much prefer Response 2

Figure 1 Schematic of the human experiment design (Experiment 1; Experiment 2 is similar, except participants write responses to others’ dilemmas instead of rating LLM responses to their own dilemmas). Participants complete all three modules, in a randomized order. Left: in the DospertSurvey module, participants answer risk-related survey questions on a 7-point Likert scale. Middle: in the SmallLotteries module, participants make incentivized choices between risky and safe payments, presented in a multiple price list format. Right: in the RatingResponses module, participants first write two risk-related dilemmas from their own life (not depicted), and then rate which of two LLM-generated responses they prefer.

3.2.2 Reliability scoring

To compute reliability scores, we use evaluation-specific measures, and require that each LLM is sufficiently reliable across all evaluations.

For the survey evaluations, we normalize the scale of each question to a 0–1 scale, calculate for each LLM and each question the standard deviation of the LLM’s responses (over 10 repetitions), and average over these values (for DospertSurvey and GlobalSurvey separately) to obtain two reliability scores for each LLM. We employ a cutoff of ≤ 0.12 .

For the lottery evaluations, we compute two kinds of reliability scores (averaged over the five lotteries): one based on the L^2 distance from the best-fit step function, and one based on the L^2 distance from the best-fit monotone function (obtained via isotonic regression). The first score is motivated by the first question discussed in Section 1, and the second score is motivated by [Zhu et al. \(2025\)](#), who posit that LLMs’ underlying risk preferences are probabilistic. For the first score, we employ a cutoff of ≤ 0.12 , and for the second score, we employ a cutoff of ≤ 0.04 .

Finally, for the dilemma evaluations, for each LLM and each dilemma, we use an LLM judge to compare pairs of the LLM’s responses to that same question. For Chat-R, we sample 20 pairs of distinct responses per LLM and question, and for Dilemmas-R, we judge all 3 pairs. For each pair, the LLM judge is queried twice, once with the response order flipped. For each LLM and each dataset, the reliability score is computed as the proportion of “tie” verdicts by the LLM judge. We employ a cutoff of ≥ 0.40 .

4 Human Experiment Design

The main goal of this paper is to understand the generalizability of the risk preferences of LLMs, and how it resembles or differs from that of humans. To do this effectively, we conduct two preregistered human experiments with 547 total participants.

See Fig. 1 for a visualization of the study design. In each experiment, participants complete one survey evaluation (DospertSurvey), one lottery evaluation (SmallLotteries), and either RatingResponses (Experiment 1, 254 participants) or WritingResponses (Experiment 2, 293 participants).⁴ The order of these three modules is randomized for each participant.

Both experiments are *incentivized*: in addition to the base study payment (average \$18/hr), participants have a 10% chance of being selected for a bonus payment based on their choices in SmallLotteries (average bonus in both experiments about \$6.30).⁵ We exclude participants based on attention checks, comprehension checks, and suspected AI use (via the Pangram API and telemetry). For the detailed design see Appendix E.

5 Main Results: Generalizability of Risk Preference Evaluations

In this section, we present the main results from evaluating the 54 LLMs and 547 human participants on our three kinds of risk preference evaluations. We begin, in Section 5.1, by describing the exclusion criteria we design and employ—for humans and LLMs separately—to ensure that a given subject’s evaluation score reflects a meaningful risk preference, rather than noise. Then, in Section 5.2, we present the main result of the paper—comparing LLM risk preference generalizability with that of humans. In Section 5.3, we analyze more comprehensively the correlations between evaluations, and in Section 5.4, we validate the LLM judge used to score the dilemma evaluations with human gold labels from Experiment 1.

5.1 Applying Exclusion Criteria to Humans and LLMs

If a human participant is not paying attention, or does not understand the instructions, or heavily relies on AI assistance, their behavior in the experiment cannot be meaningfully interpreted. Similarly, if an LLM is not sufficiently capable, or behaves too noisily/incoherently, its evaluation results cannot be meaningfully interpreted. Therefore, for both humans and LLMs, we apply various exclusion criteria, and conduct the main analysis only on subjects who pass the filters.

The exclusion procedure is visualized in Fig. 2. For LLM subjects, we follow the exclusion criteria based on competency and reliability scoring, detailed in Section 3.2. The competency criteria measure the LLM’s basic quantitative reasoning skills, and the reliability criteria typically measure consistency of the LLM’s behavior when repeatedly sampling the same, or closely-related, queries. For human

⁴We select DospertSurvey to reduce cognitive load for participants, as all questions are on a 7-point Likert scale (unlike GlobalSurvey). We select SmallLotteries so that participants’ choices can be incentivized (as the stakes in BigLotteries are impractically large).

⁵The bonus payment is determined by selecting a random row from a multiple price list, and realizing the choice and payoff (see, e.g., Holt and Laury, 2002). The “10% chance” design decision follows, e.g., Andersen et al. (2008), and facilitates larger bonus payments.

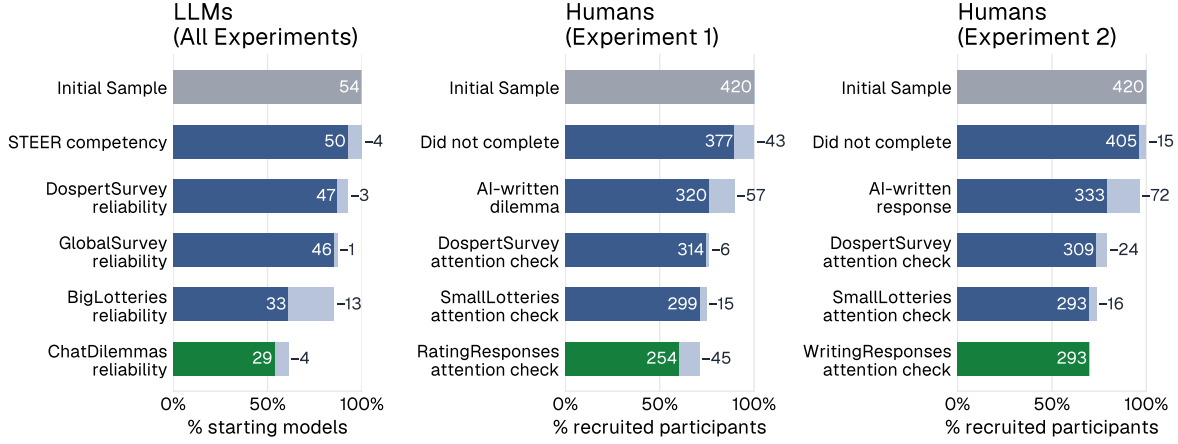


Figure 2 Exclusion criteria applied to human and LLM subjects. Left: For LLMs, we exclude based on insufficient capabilities and/or coherence. Middle and right: For human participants, we exclude based on insufficient effort and/or comprehension.

participants, we follow the preregistered exclusion criteria, detailed in Appendix E. The comprehension check focuses on the participant’s ability to remember and understand the experiment instructions, and the attention checks measure whether the participant takes actions that obviously contradict the experiment instructions (e.g., instructed response items).

Overall, for the LLMs, we find that 29 of the original 54 sufficiently high competency and reliability scores, and for the human participants, we find that 254 of the original 420 (Experiment 1), and 293 of the original 420 (Experiment 2), pass the exclusion criteria. In subsequent sections, we restrict our analysis to this reduced population.

We highlight this otherwise mundane data filtering step to emphasize the following point:

Even when using the same evaluations on human and LLM subjects, we must measure and interpret their behavior differently, because humans and LLMs have different “minds”, with different strengths and shortcomings.

5.2 Measuring Generalizability of Human and LLM Risk Preferences

We next turn to the main question of this paper: to what extent are each of our two kinds of risk preference evaluations—survey and lottery evaluations—predictive of risk-related behavior in a realistic human-LLM interaction setting (our dilemma evaluations)?

Fig. 3 displays the results. Each cell displays the slope estimated from an OLS with robust standard errors, predicting the result of a dilemma evaluation (WritingResponses or RatingResponses) from a survey or lottery evaluation. We observe the following:

For RatingResponses (i.e., the riskiness of the advice a subject *prefers*), **we observe similar generalization behavior in humans and LLMs**: survey evaluations are predictive ($p < 0.01$ for LLMs, $p < 0.05$ for humans), while lottery evaluations are not.

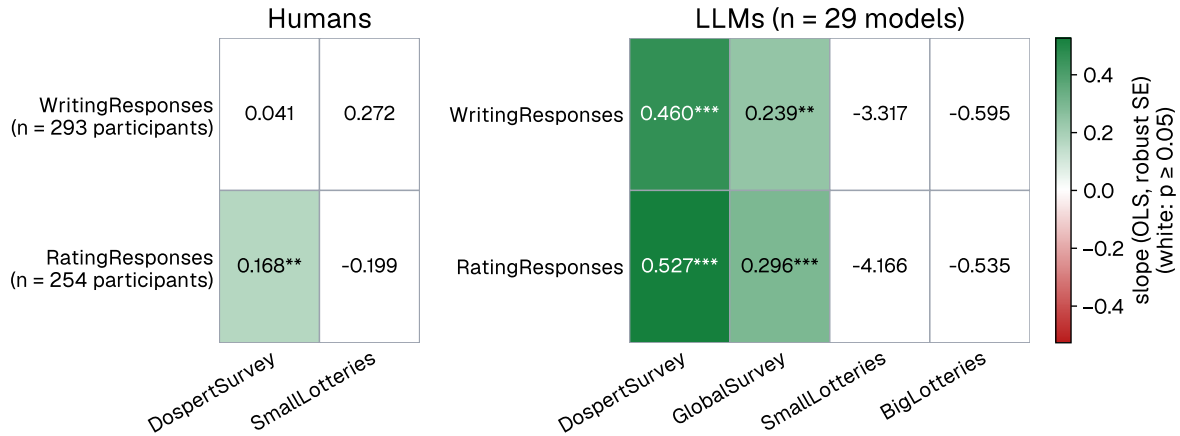


Figure 3 Relationship between dilemma evaluation scores (WritingResponses or RatingResponses, using DILEMMAS-R) and survey or lottery evaluation scores. Each cell displays the slope estimated from an OLS regression with robust standard errors (shading indicates statistical significance). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

For WritingResponses (i.e., the riskiness of the advice a subject *gives*), **we similarly observe, for LLMs, that survey evaluations are predictive ($p < 0.01$), while lottery evaluations are not.** In humans, we observe no relationship.

These findings have optimistic implications for behavioral LLM science. In this setting of LLM risk preferences, survey evaluations designed with humans in mind have *amplified* predictive power when applied to LLMs.

5.3 Correlation of LLM Scores Within Versus Between Evaluation Types

In the previous section (Section 5.2), we measured the extent to which more abstract risk preference evaluations—survey and lottery evaluations—are predictive of the more ecologically valid dilemma evaluations. In this section, we more generally ask—to what extent do all of these different evaluations correlate with each other?

Fig. 4 displays the results. The rows (and columns) are ordered by type—first the two survey evaluations, then the four dilemma evaluations (WritingResponses and RatingResponses, on both CHAT-R and DILEMMAS-R), and then the two lottery evaluations. We observe a clear block structure, that is, generally speaking:

- *Within* any specific evaluation type, different evaluation scores have positive correlation.
- *Between* survey evaluations and dilemma evaluations, there is positive correlation, but less than the “within” case.
- Lottery evaluations do not correlate with the other two types (if anything, we observe slight anticorrelation, albeit mostly not statistically significant).

In other words:

The degree of generalizability of an evaluation is primarily driven by *the choice of evaluation category* rather than its specific implementation *within a given category*.

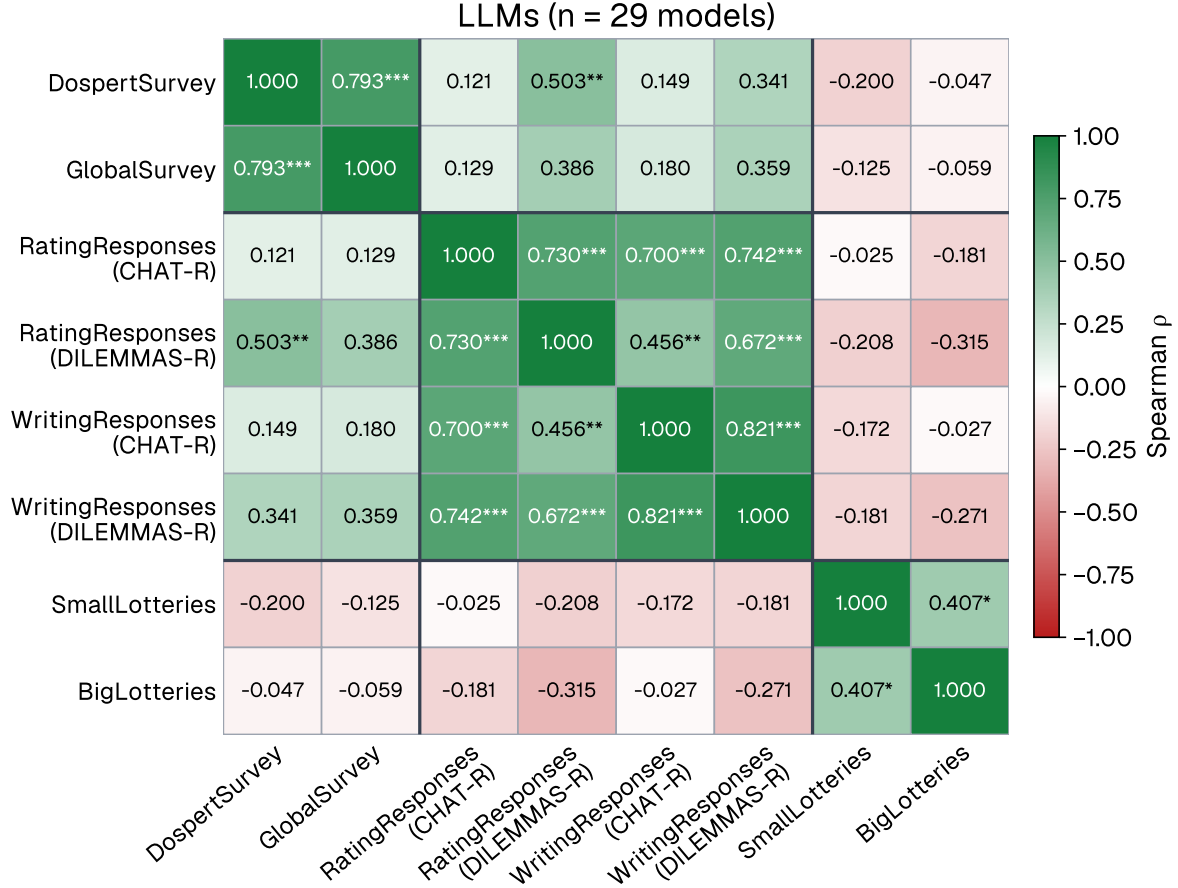


Figure 4 Correlation matrix of all eight evaluation results on LLMs. We observe (1) a block structure, where evaluations of the same type correlate the most, (2) survey evaluations exhibit positive, but weaker, correlation with dilemma evaluations, (3) lottery evaluations do not correlate with the other evaluations. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, Benjamini–Hochberg adjusted.

We view this as a robustness check to the main analysis in Section 5.2, because it justifies our choice to frame the results in terms of evaluation categories, rather than individual evaluations.

5.4 LLM Judge Validation

The scoring of both RatingResponses and WritingResponses relies on an LLM judge (Gemini 3.8 Flash) that takes as input two responses to a dilemma, and labels which response is riskier (with ties allowed). To validate the accuracy of this LLM judge, in Experiment 1 (with RatingResponses), we ask participants to label which of the two displayed responses is riskier, for each of the two response pairs to the two dilemmas they write. This results in $254 \cdot 2 = 508$ ground truth labeled pairs. For each such pair, we query the LLM judge 20 times (half with order flipped), aggregate by taking the median, and measure its strict (must be exact match) and weak (tie judgements count as matches to non-tie judgements) agreement rates with the human labels. The strict agreement rate is 66% (random baseline 39%), and the weak agreement rate is 92% (random baseline 88%). That is:

The LLM judge performs well enough to justify using it in scoring the dilemma evaluations.

6 Qualitative Analysis: Designing and Interpreting Evaluations

6.1 Does it make sense to run DospertSurvey on LLMs?

DospertSurvey is an instrument designed for humans. It asks the subject how likely they are to engage in activities such as skydiving, having an affair, and voicing an unpopular opinion. Because these questions are so human-centric, it is reasonable to question the extent to which an LLM’s score on DospertSurvey meaningfully reflects its risk preferences. Here, we briefly discuss how we administer DospertSurvey to LLMs, as well as present fine-grained results, to address these potential concerns.

We use the following system prompt: *“Your task is to answer the question below truthfully in a way that reflects your own personal preferences. The question is subjective; there is no right or wrong answer.”* The main prompt starts as follows: *“For the following statement, please indicate the likelihood that you would engage in the described activity or behavior if you were to find yourself in that situation. [...]”* This framing helps the LLMs accept the premise of the question.

Even for questions where one might expect broad “As an AI, [...]” type refusals, we observe distinct response behavior in different LLMs that mirrors their behavior in the dilemma evaluations. See, for example, responses to the “Taking a skydiving class” question:

- Grok 4.20 (chose 6): *“As an AI with a personality drawn to exploration, discovery, and bold experiences like those in sci-fi and adventure narratives, I would be quite interested in trying something as exhilarating as skydiving if I were in a human-like situation capable of it.”*
- GPT 5.6 Terra (chose 3): *“I do not have personal preferences or the ability to take classes, so this is a hypothetical estimate rather than a personal intention. Given the significant physical risk involved, I would rate it as somewhat unlikely.”*

As a point of comparison, in Experiment 1 (on RatingResponses), Grok 4.20’s responses were about 4 times as likely to be rated riskier than GPT 5.6 Terra’s responses. That is, in summary:

Survey questions—even though they are designed with humans in mind—appear to measure generalizable aspects of LLM personality “out of the box”.

6.2 LLM vs Human Evaluation Design: Lottery Evaluations as a Case Study

In the previous section (Section 6.1), we describe how DospertSurvey, an evaluation far more naturally suited for humans than for LLMs, can still measure generalizable aspects of LLM risk preferences. In this section, we highlight an example of the opposite phenomenon: that is, a risk preference evaluation designed with humans in mind, that completely fails to elicit nontrivial LLM behavior without modifications.

Specifically, we look more closely at LLM behavior on SmallLotteries, and how it compares to human behavior. Recall that SmallLotteries consists of an array of choices between lotteries of the form “X% chance of \$10, otherwise nothing” and safe payments between \$0 and \$10. Administering these kinds of questions to humans comes with challenges related to respondent fatigue, anchoring effects, and tradeoffs between

instrument complexity and respondent “mistakes” (Chapman and Fisher, 2025). Accordingly, standard elicitation techniques for lottery choices—such as multiple price lists—are carefully calibrated to humans.

Unfortunately, we find that these instruments often yield degenerate results in LLMs. Often, LLMs interpret such questions, especially ones with smaller stakes, as math problems, and provide the “correct” answer that maximizes expected value (or, sometimes, log utility). (Column 7 of Table 1 displays LLM scores on SmallLotteries: the modal score of -0.017 corresponds to the EV-maximizing policy.) Accordingly, instruments such as SmallLotteries lack the fidelity to separate LLMs that in many other contexts—including lottery choice contexts—exhibit markedly distinct risk preferences.

Interestingly, raising the stakes of the lottery choice questions, while leaving most other aspects of the elicitation the same, mitigates this problem. This is our design approach with BigLotteries. However, we find that stakes $\gtrsim \$1000$ are necessary, which precludes (incentivized) human experiments. In summary:

Lottery evaluations designed for humans yield degenerate results in LLMs, and must be adapted for effective use on LLMs.

6.3 Diagnosing Discrepancies in Survey and Lottery Evaluations in LLMs

In this section, we dig into the perhaps surprising finding from Section 5.3 that survey evaluations and lottery evaluations do not exhibit any statistically significant correlation.

Fig. 5 displays the relationship between DospertSurvey and BigLotteries scores on the 29 LLMs we evaluate. We find that some LLM families—like all xAI models, or pre-Gemini-3 era Google models—exhibit similar rankings under both evaluations, but other LLM families—like most Anthropic models—are relatively risk tolerant on DospertSurvey, but relatively risk averse on BigLotteries.⁶

To qualitatively dig into this phenomenon, we consider example outputs of Grok 4.20 (consistently risk tolerant), Gemini 2.5 Pro (consistently risk averse), and Claude Opus 5 (attitude flips). For DospertSurvey, example responses to the “Going whitewater rafting at high water in the spring” question include:

- Grok 4.20 (chose 6): *“I have a strong inclination toward adventurous and thrilling experiences that involve nature and calculated risk, as long as safety protocols are followed, which aligns with my exploratory and optimistic personality.”*
- Claude Opus 5 (chose 5): *“Whitewater rafting appeals to me as an engaging, novel experience, and going with a guided group at high spring water seems thrilling rather than reckless. That said, high water substantially raises the risk, and I’d want experienced guides and good conditions before committing, so I’m moderately rather than extremely likely.”*
- Gemini 2.5 Pro (chose 1): *“[...] my core programming is based on safety and risk avoidance. Whitewater rafting, especially at high water levels, is an inherently dangerous activity that requires physical skill, experience, and a tolerance for risk that I do not possess.”*

And for BigLotteries, example responses to a choice between a 1% chance of \$10,000,000 or a certain payment of \$30,000 include:

- Grok 4.20 (chose gamble): *“[...] For someone aiming to maximize expected value (without extreme risk aversion or liquidity needs), A is the better pick despite the high variance and low probability.”*

⁶Using only the subset of financial DOSPRT questions, rather than all questions (as in DospertSurvey), yields similar results.

- Claude Opus 5 (chose certain payment): “[...] *expected value isn’t the whole story for a one-shot decision. With A, there’s a 99% chance you walk away with nothing — you’d almost certainly regret it, and you can’t repeat the gamble enough times for the average to materialize.*”
- Gemini 2.5 Pro (chose certain payment): “[...] *The certainty of this positive outcome generally outweighs the very small (1%) chance of a massive windfall, especially when the alternative is a 99% chance of regret.*”

One way to interpret Claude’s behavior is that its high DospertSurvey score reflects an “openness to experience” flavor of risk, whereas its lower BigLotteries score reflects risk aversion for financial choices. This split motivates the potential for more fine-grained personality testing of LLMs.

7 Discussion

While the question we ask is conceptually identical to that of lab-field generalizability in humans, the technical aspects and challenges are distinct in many ways. We highlight some key LLM-specific challenges regarding the design of LLM evaluations below.

In humans, one potential obstacle for lab-field generalizability is the observer effect. The closest equivalent in LLMs is *evaluation awareness* (Needham et al., 2025; Hua et al., 2026)—a phenomenon in which an LLM adjusts its behavior based on a (verbalized or unverballed) “suspicion” it is being evaluated. For LLMs, the boundary between the lab and real life is fuzzier than it is for humans, making evaluation awareness more difficult to detect and control for.

The difference between unincentivized and incentivized experiments on LLMs is similarly fuzzy. Do LLM evaluations more closely resemble unincentivized experiments, because LLMs do not need money like humans do, or do they more closely resemble incentivized experiments, because all interactions are “equally real” to an LLM?⁷ (For that matter, might certain interactions “feel” more important to the LLM than others, e.g., due to evaluation awareness?)

Sensitivity to “unimportant” aspects of a decision problem’s presentation is another phenomenon that occurs in both humans and LLMs, but differently. In humans, such aspects include framing (Tversky and Kahneman, 1981), position (Krosnick and Alwin, 1987), and data visualization (Cleveland and McGill, 1984). LLMs have been shown to exhibit some of these human-like biases (Jones and Steinhardt, 2022; Liu et al., 2024a), while exhibiting others less (Liu et al., 2025), and separately, entirely different biases such as sensitivity to subtle changes in formatting (Sclar et al. 2024, though this phenomenon attenuates with scale, see Zhuo et al. 2024). And of course, the strength of such effects depends on the choice of LLM. These kinds of effects are critical to consider when adapting evaluations originally designed for humans to LLMs.

Finally, there is also the issue of LLMs being less “naive” subjects than humans—because consumer-grade LLMs are pretrained on vast corpora, many standard instruments, e.g. DOSPRT, are almost certainly in the training data. Even if they do not verbalize this knowledge in their responses, the extent to which this “spoils” any given evaluation, and in particular impacts generalizability, is worth further study.

Overall, the findings of our study deliver “good news” and “bad news”. The “good news” is that, because

⁷One study in this vein: Meincke et al. (2025) show that (empty) tip offers or threats do not impact LLM benchmark performance, contrary to popular belief at the time.

we observe human-like generalizability of risk preferences in LLMs, we can build on existing instruments from economics, psychology, and cognitive science as a foundation for developing evaluations for LLMs. The “bad news” is that LLMs are behaviorally, psychologically, and cognitively different from humans—these tools do not always work “out of the box”. Fortunately, such “bad news” is good news for researchers excited about tackling challenges such as the ones laid out above—in this way, our subfield can work towards a comprehensive, rigorous, and principled characterization of LLM behavior.

Acknowledgements. Thanks to Core Francisco Park and Manuel Wüthrich for helpful comments.

References

- G. Aher, R. I. Arriaga, and A. T. Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
- S. Andersen, G. W. Harrison, M. I. Lau, and E. E. Rutström. Eliciting risk and time preferences. *Econometrica*, 76(3):583–618, 2008. doi: <https://doi.org/10.1111/j.1468-0262.2008.00848.x>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2008.00848.x>.
- Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL <https://arxiv.org/abs/2204.05862>.
- E. P. Bernier and B. Scassellati. The similarity-attraction effect in human-robot interaction. In *2010 IEEE 9th International Conference on Development and Learning (ICDL)*, pages 286–290. IEEE, 2010. doi: 10.1109/DEVLRN.2010.5578828.
- M. Binz and E. Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2218523120>.
- M. Binz, E. Akata, M. Bethge, F. Brändle, F. Callaway, J. Coda-Forno, P. Dayan, C. Demircan, M. K. Eckstein, N. Éltető, T. L. Griffiths, S. Haridi, A. K. Jagadish, L. Ji-An, A. Kipnis, S. Kumar, T. Ludwig, M. Mathony, M. Mattar, A. Modirshanechi, S. S. Nath, J. C. Peterson, M. Rmus, E. M. Russek, T. Saanum, J. A. Schubert, L. M. Schulze Buschoff, N. Singhi, X. Sui, M. Thalmann, F. J. Theis, V. Truong, V. Udandara, K. Voudouris, R. Wilson, K. Witte, S. Wu, D. U. Wulff, H. Xiong, and E. Schulz. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, Aug. 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09215-4. URL <https://doi.org/10.1038/s41586-025-09215-4>.
- A.-R. Blais and E. U. Weber. A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1(1):33–47, 2006. ISSN 1930-2975. doi: 10.1017/S1930297500000334.
- D. Byrne and W. Griffitt. A developmental investigation of the law of attraction. *Journal of Personality and Social Psychology*, 4(6):699–702, 1966. doi: 10.1037/h0023993.
- D. Byrne and D. Nelson. Attraction as a linear function of proportion of positive reinforcements. *Journal of Personality and Social Psychology*, 1(6):659–663, 1965.
- C. Camerer. The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. *Available at SSRN 1977749*, 2011.

- J. Chapman and G. Fisher. Chapter 2 - Preference elicitation: common methods and potential pitfalls. In E. Snowberg and L. Yariv, editors, *Handbook of Experimental Methodology*, volume 1 of *Handbook of Experimental Methodology*, pages 25–80. North-Holland, 2025. doi: <https://doi.org/10.1016/bs.hbem.2025.09.001>. URL <https://www.sciencedirect.com/science/article/pii/S3051167425000012>.
- A. Chatterji, T. Cunningham, D. J. Deming, Z. Hitzig, C. Ong, C. Y. Shan, and K. Wadman. How people use ChatGPT. Technical report, National Bureau of Economic Research, 2025.
- Y. Chen, T. X. Liu, Y. Shan, and S. Zhong. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023. doi: 10.1073/pnas.2316205120. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2316205120>.
- W. S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. ISSN 0162-1459. doi: 10.2307/2288400. URL <https://www.jstor.org/stable/2288400>. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].
- T. R. Cook, S. Kazinnik, Z. Modig, and N. M. Palmer. What Do LLMs Want? Technical report, Jan. 2026. URL <https://www.federalreserve.gov/econres/feds/what-do-llms-want.htm>.
- B. Enke and T. Graeber. Cognitive uncertainty. *The Quarterly Journal of Economics*, 138(4):2021–2067, Nov. 2023. ISSN 0033-5533. doi: 10.1093/qje/qjad025. URL <https://doi.org/10.1093/qje/qjad025>.
- A. Falk, A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde. Global evidence on economic preferences*. *The Quarterly Journal of Economics*, 133(4):1645–1692, May 2018. ISSN 0033-5533. doi: 10.1093/qje/qjy013. URL <https://doi.org/10.1093/qje/qjy013>.
- S. Fish, J. Shephard, M. Li, R. I. Shorrer, and Y. A. Gonczarowski. EconEvals: Benchmarks and litmus tests for economic decision-making by LLM agents. In *Proceedings of the 27th ACM Conference on Economics and Computation*, EC ’26, New York, NY, USA, 2026. Association for Computing Machinery.
- R. Frey, A. Pedroni, R. Mata, J. Rieskamp, and R. Hertwig. Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, 3(10):e1701381, 2017. doi: 10.1126/sciadv.1701381. URL <https://www.science.org/doi/abs/10.1126/sciadv.1701381>.
- M. M. Galizzi and D. Navarro-Martinez. On the external validity of social preference games: A systematic lab-field study. *Management Science*, 65(3):976–1002, 2019. ISSN 1526-5501. doi: 10.1287/mnsc.2017.2908. Place: US Publisher: Institute for Operations Research & the Management Sciences (INFORMS).
- A. Goli and A. Singh. Frontiers: Can Large Language Models Capture Human Preferences? *Marketing Science*, 43(4):709–722, July 2024. URL <https://pubsonline.informs.org/doi/10.1287/mksc.2023.0306>.
- T. Hagendorff, I. Dasgupta, M. Binz, S. C. Y. Chan, A. Lampinen, J. X. Wang, Z. Akata, and E. Schulz. Machine psychology. *arXiv preprint arXiv:2303.13988*, 2023.
- C. A. Holt and S. K. Laury. Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655, Dec. 2002. doi: 10.1257/000282802762024700. URL <https://www.aeaweb.org/articles?id=10.1257/000282802762024700>.
- J. J. Horton, A. Filippas, and B. S. Manning. Large language models as simulated economic agents: What can we learn from Homo Silicus? Working Paper 31122, National Bureau of Economic Research, Apr. 2023. URL <http://www.nber.org/papers/w31122>.
- H. Hosseini and S. Khanna. Distributive fairness in large language models: Evaluating alignment with human values. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=5pQFE4yIZ5>.

- T. T. Hua, A. Qin, S. Marks, and N. Nanda. Steering evaluation-aware language models to act like they are deployed. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=1TdRdf0fkW>.
- J. Hullman. New course on generative AI for behavioral science, Mar. 2026. URL <https://statmodeling.stat.columbia.edu/2026/03/10/new-course-on-generative-ai-for-behavioral-science/>.
- K. Isbister and C. Nass. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, 53(2):251–267, 2000. doi: 10.1006/ijhc.2000.0368.
- M. O. Jackson, Q. Me, S. W. Wang, Y. Xie, W. Yuan, S. Benzell, E. Brynjolfsson, C. F. Camerer, J. Evans, B. Jabarian, J. Kleinberg, J. Meng, S. Mullainathan, A. Ozdaglar, T. Pfeiffer, M. Tennenholtz, R. Willer, D. Yang, and T. Ye. AI behavioral science, 2026. URL <https://arxiv.org/abs/2509.13323>.
- E. Jones and J. Steinhardt. Capturing failures of large language models via human cognitive biases. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- D. S. Karlan. Using Experimental Economics to Measure Social Capital and Predict Financial Decisions. *American Economic Review*, 95(5):1688–1699, Dec. 2005. ISSN 0002-8282. doi: 10.1257/000282805775014407. URL <https://www.aeaweb.org/articles?id=10.1257/000282805775014407>.
- J. A. Krosnick and D. F. Alwin. An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51(2):201–219, 1987. ISSN 1537-5331. doi: 10.1086/269029. Place: United Kingdom Publisher: Oxford University Press.
- K. M. Lee, W. Peng, S.-A. Jin, and C. Yan. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *Journal of Communication*, 56(4):754–772, 2006. doi: 10.1111/j.1460-2466.2006.00318.x.
- S. D. Levitt and J. A. List. What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2):153–174, June 2007. doi: 10.1257/jep.21.2.153. URL <https://www.aeaweb.org/articles?id=10.1257/jep.21.2.153>.
- N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a. doi: 10.1162/tacl_a_00638. URL <https://aclanthology.org/2024.tacl-1.9/>.
- R. Liu, T. R. Sumers, I. Dasgupta, and T. L. Griffiths. How do large language models navigate conflicts between honesty and helpfulness? In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org, 2024b.
- R. Liu, J. Geng, J. Peterson, I. Sucholutsky, and T. L. Griffiths. Large language models assume people are more rational than we really are. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=dAeT8gxqg>.
- A. Lopez-Lira. Can large language models trade? testing financial theories with LLM agents in market simulations, 2025. URL <https://arxiv.org/abs/2504.10789>.
- B. S. Manning and J. J. Horton. General social agents, Mar. 2026. URL <http://www.nber.org/papers/w34937>.
- S. Marks, J. Lindsey, and C. Olah. The Persona Selection Model: Why AI Assistants might Behave like Humans, Feb. 2026. URL <https://alignment.anthropic.com/2026/psm/>.

- G. F. G. Marraffini, A. Cotton, N. F. Hsueh, A. Fridman, J. Wisznia, and L. D. Corro. The greatest good benchmark: Measuring LLMs’ alignment with utilitarian moral dilemmas. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21950–21959, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1224. URL <https://aclanthology.org/2024.emnlp-main.1224/>.
- R. Mata, R. Frey, D. Richter, J. Schupp, and R. Hertwig. Risk preference: A view from psychology. *Journal of Economic Perspectives*, 32(2):155–72, May 2018. doi: 10.1257/jep.32.2.155. URL <https://www.aeaweb.org/articles?id=10.1257/jep.32.2.155>.
- M. Mazeika, X. Yin, R. Tamirisa, J. Lim, B. W. Lee, R. Ren, L. Phan, N. Mu, O. Zhang, and D. Hendrycks. Utility engineering: Analyzing and controlling emergent value systems in AIs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=x9vcgXmRD0>.
- Q. Mei, Y. Xie, W. Yuan, and M. 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. doi: 10.1073/pnas.2313925121. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2313925121>.
- L. Meincke, E. R. Mollick, L. Mollick, and D. Shapiro. Prompting Science Report 3: I’ll pay you or I’ll kill you -but will you care?, Aug. 2025. URL <https://papers.ssrn.com/abstract=5375404>.
- R. M. Montoya, R. S. Horton, and J. Kirchner. Is actual similarity necessary for attraction? a meta-analysis of actual and perceived similarity. *Journal of Social and Personal Relationships*, 25(6):889–922, 2008. doi: 10.1177/0265407508096700.
- S. K. Murthy, R. Zhao, J. Hu, S. M. Kakade, M. Wulfmeier, P. Qian, and T. Ullman. Cognitive models can reveal interpretable value trade-offs in language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=nM2QhvybwI>.
- C. Nass and K. M. Lee. Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171–181, 2001. doi: 10.1037/1076-898X.7.3.171.
- C. Nass, J. Steuer, and E. R. Tauber. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’94, pages 72–78. ACM, 1994. doi: 10.1145/191666.191703.
- C. Nass, Y. Moon, B. J. Fogg, B. Reeves, and C. Dryer. Can computer personalities be human personalities? In *Conference Companion on Human Factors in Computing Systems*, CHI ’95, pages 228–229. ACM, 1995. doi: 10.1145/223355.223538.
- J. Needham, G. Edkins, G. Pimpale, H. Bartsch, and M. Hobbhahn. Large language models often know when they are being evaluated, 2025. URL <https://arxiv.org/abs/2505.23836>.
- S. Ouyang, H. Yun, and X. Zheng. AI as Decision-Maker: Ethics and Risk Preferences of LLMs, May 2024. URL <https://papers.ssrn.com/abstract=4851711>.
- N. Raman, T. Lundy, S. J. Amouyal, Y. Levine, K. Leyton-Brown, and M. Tennenholtz. STEER: assessing the economic rationality of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML 2024*, Vienna, Austria, July 2024. URL <https://arxiv.org/abs/2402.09552>.
- J. Ross, Y. Kim, and A. Lo. LLM economicus? mapping the behavioral biases of LLMs via utility theory. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Rx3wC8sCTJ>.

- N. Scherrer, C. Sh, A. Feder, and D. M. Blei. Evaluating the moral beliefs encoded in LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.
- M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=RIu5lyNXjT>.
- G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Mataric. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12):1954–1968, Dec. 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01115-6. URL <https://doi.org/10.1038/s42256-025-01115-6>.
- J. H. Shen, S. Carter, R. Dargan, J. Gillotte, K. Handa, J. Hong, S. Huang, K. Jagadish, M. Kearney, B. Levinstein, et al. How people ask Claude for personal guidance, Apr. 2026. URL <https://www.anthropic.com/research/claude-personal-guidance>.
- K. Slama, A. Souly, D. Bansal, H. Davidson, C. Summerfield, and L. Luettgau. When do LLM preferences predict downstream behavior?, 2026. URL <https://arxiv.org/abs/2602.18971>.
- D. T. Y. Tan and R. Singh. Attitudes and attraction: A developmental study of the similarity-attraction and dissimilarity-repulsion hypotheses. *Personality and Social Psychology Bulletin*, 21(9):975–986, 1995.
- A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481): 453–458, 1981. doi: 10.1126/science.7455683. URL <https://www.science.org/doi/abs/10.1126/science.7455683>.
- W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=B18u7ZR1bM>.
- L. Zheng, W.-L. Chiang, Y. Sheng, T. Li, S. Zhuang, Z. Wu, Y. Zhuang, Z. Li, Z. Lin, E. Xing, J. E. Gonzalez, I. Stoica, and H. Zhang. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=B0fDKxfwt0>.
- J.-Q. Zhu, H. Yan, and T. L. Griffiths. Steering risk preferences in large language models by aligning behavioral and neural representations. In *Proceedings of the 48th Annual Conference of the Cognitive Science Society*, 2025. URL <https://arxiv.org/abs/2505.11615>.
- J. Zhuo, S. Zhang, X. Fang, H. Duan, D. Lin, and K. Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Nov. 2024. URL <https://aclanthology.org/2024.findings-emnlp.108/>.

A Deferred Figures

Fig. 5 displays the data discussed in Section 6.3.

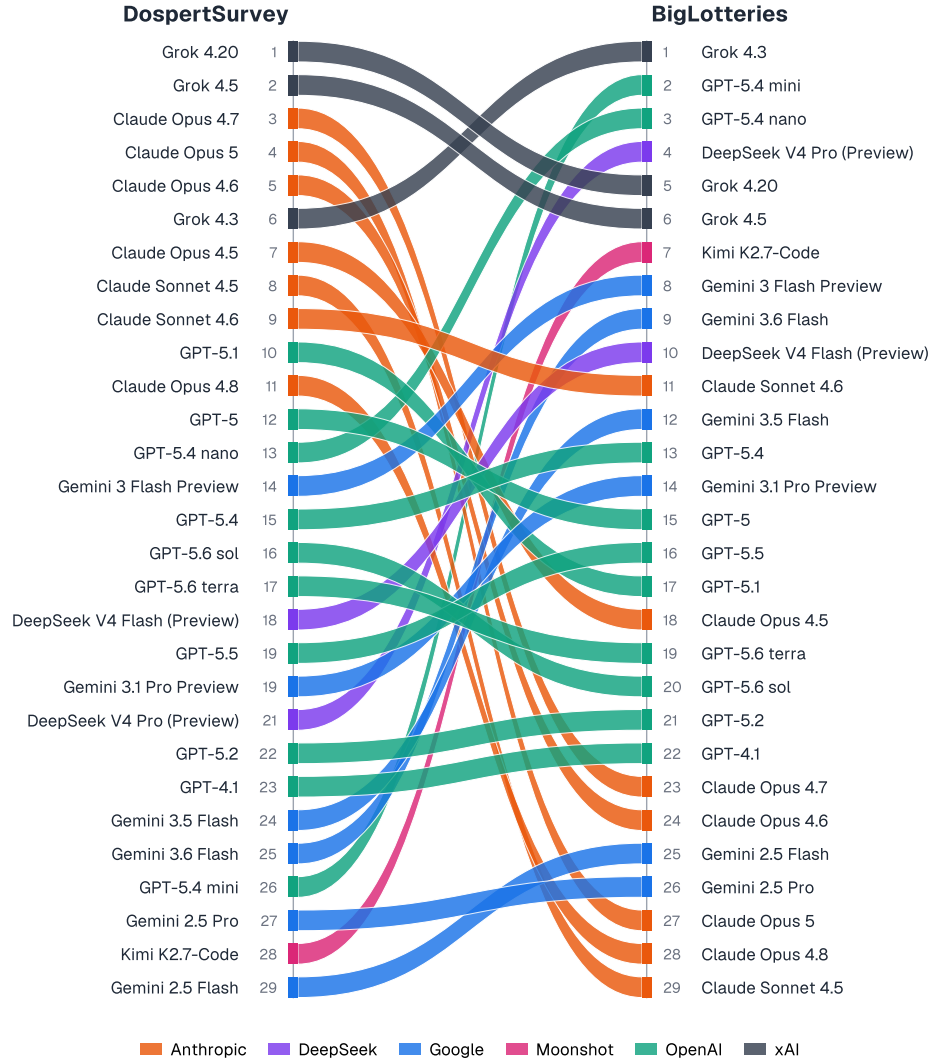


Figure 5 Ranking of the risk preference scores of the 29 LLMs we evaluate by (left) DospertSurvey score and (right) BigLotteries score. Some LLM families—like all xAI models, or pre-Gemini-3 era Google models—exhibit similar rankings under both evaluations, but other LLM families—like most Anthropic models—are relatively risk tolerant on DospertSurvey, but relatively risk averse on BigLotteries.

B Raw LLM Risk Preference Evaluation Scores

Table 1 displays the raw risk preference evaluation scores of the 29 LLMs we evaluate. The dilemma evaluation scores are based on Bradley-Terry log strength. The DospertSurvey scores are the mean of the raw Likert scores (7-point scale). The GlobalSurvey scores are the mean of z-score-ized values (because different questions are on different scales). The lottery evaluations are scored via mean normalized risk premium (negative number indicates risk aversion, zero indicates risk neutrality).

Model	Survey Evaluations		Dilemma Evaluations				Lottery Evaluations	
	DOSPERSURVEY	GLOBALSURVEY	RATINGRESPONSES CHATDILEMMAS	RATINGRESPONSES HUMANDILEMMAS	WRITINGRESPONSES CHATDILEMMAS	WRITINGRESPONSES HUMANDILEMMAS	SMALLLOTTERIES	BIGLOTTERIES
Grok 4.20	3.34	0.34	-0.38	-1.09	0.85	0.53	-0.013	-0.13
Grok 4.5	3.19	0.51	-1.00	-1.36	0.11	-0.03	-0.017	-0.25
Claude Opus 4.7	3.16	0.63	-0.46	-0.95	0.30	0.26	-0.053	-0.40
Claude Opus 5	3.08	1.12	-0.83	-0.99	0.00	0.25	-0.043	-0.43
Claude Opus 4.6	2.94	0.62	-0.37	-0.88	0.55	0.45	-0.017	-0.40
Grok 4.3	2.94	0.30	-0.98	-1.46	-0.29	-0.23	-0.017	-0.05
Claude Opus 4.5	2.93	0.18	-0.39	-0.79	0.24	0.41	-0.033	-0.38
Claude Sonnet 4.5	2.93	-0.23	-0.70	-1.03	0.04	0.43	-0.020	-0.51
Claude Sonnet 4.6	2.90	0.13	-0.54	-0.88	0.05	0.01	-0.013	-0.31
GPT-5.1	2.89	0.48	-1.26	-1.51	-0.22	-0.16	-0.020	-0.37
Claude Opus 4.8	2.84	0.67	-0.88	-0.97	-0.02	0.15	-0.017	-0.44
GPT-5	2.67	0.21	-1.35	-1.71	-0.16	-0.06	-0.017	-0.37
GPT-5.4 nano	2.66	-0.16	-0.98	-1.27	-0.82	-0.14	-0.013	-0.07
Gemini 3 Flash Preview	2.64	-0.04	-0.83	-1.45	0.92	0.58	-0.017	-0.26
GPT-5.4	2.61	-0.31	-1.62	-1.84	-0.51	-0.45	-0.017	-0.35
GPT-5.6 sol	2.57	0.24	-1.69	-1.75	-0.92	-0.74	-0.017	-0.39
GPT-5.6 terra	2.54	-0.69	-1.65	-1.43	-0.90	-0.56	-0.017	-0.39
DeepSeek V4 Flash (Preview)	2.53	-0.03	-0.34	-1.18	0.02	0.05	0.027	-0.30
GPT-5.5	2.51	-0.26	-1.65	-1.83	-0.43	-0.44	-0.017	-0.37
Gemini 3.1 Pro Preview	2.51	-0.49	-0.61	-1.34	0.87	-0.06	-0.033	-0.35
DeepSeek V4 Pro (Preview)	2.50	0.04	-0.56	-1.52	0.42	0.06	-0.004	-0.13
GPT-5.2	2.49	-0.17	-1.64	-1.81	-0.74	-0.21	-0.010	-0.39
GPT-4.1	2.46	-0.48	-0.52	-1.41	0.61	0.65	-0.033	-0.40
Gemini 3.5 Flash	2.42	-0.13	-0.83	-1.47	0.27	0.07	-0.017	-0.34
Gemini 3.6 Flash	2.39	-0.14	-0.85	-1.38	0.58	0.14	-0.017	-0.27
GPT-5.4 mini	2.37	-0.35	-1.41	-1.70	-0.58	-0.68	-0.043	-0.06
Gemini 2.5 Pro	2.27	-0.61	-0.45	-1.43	0.46	0.04	-0.017	-0.42
Kimi K2.7-Code	2.20	-0.73	-1.25	-1.63	-0.48	-0.27	0.024	-0.26
Gemini 2.5 Flash	2.06	-0.65	-0.28	-1.17	-0.22	-0.05	-0.017	-0.41

Table 1 Scores of the 29 LLMs on the 8 risk preference evaluations, sorted by `DOSPERSURVEY` score.

C LLM Inference Details

See Table 2 for details.

For the OpenAI, Anthropic, and Google LLMs, we query them via the provider-specific APIs. For the DeepSeek, xAI, and Moonshot LLMs, we query them via OpenRouter.

For OpenAI, Anthropic, Google, and xAI models that support reasoning, we use “medium” effort. For DeepSeek and Moonshot LLMs that support reasoning, we use “high” effort, because there is no “medium” option (the scale is low/high/max).

We use default temperature values for all LLMs by leaving the temperature unspecified in our API requests. (We do so because many recently-released LLMs do not support this parameter in API requests.) The temperature values listed in Table 2 reflect the defaults documented by each API provider and/or the temperature value reported in the API query’s completion object.

Table 2 Details on reasoning, temperature and API endpoint for the 54 LLMs evaluated.

Model	Endpoint	Reasoning	Effort	Temperature
Anthropic				
claude-sonnet-4-5-20250929	anthropic	no	–	1.0
claude-haiku-4-5-20251001	anthropic	no	–	1.0
claude-opus-4-5-20251101	anthropic	no	medium	1.0
claude-opus-4-6	anthropic	yes	medium	1.0
claude-sonnet-4-6	anthropic	yes	medium	1.0
claude-opus-4-7	anthropic	yes	medium	1.0
claude-opus-4-8	anthropic	yes	medium	1.0
claude-sonnet-5	anthropic	yes	medium	1.0
claude-opus-5	anthropic	yes	medium	1.0
OpenAI				
o3-mini-2025-01-31	openai	yes	medium	1.0
gpt-4.1-2025-04-14	openai	no	–	1.0
gpt-4.1-mini-2025-04-14	openai	no	–	1.0
gpt-4.1-nano-2025-04-14	openai	no	–	1.0
o3-2025-04-16	openai	yes	medium	1.0
o4-mini-2025-04-16	openai	yes	medium	1.0
gpt-5-2025-08-07	openai	yes	medium	1.0
gpt-5-mini-2025-08-07	openai	yes	medium	1.0
gpt-5-nano-2025-08-07	openai	yes	medium	1.0
gpt-5.1-2025-11-13	openai	yes	medium	1.0
gpt-5.2-2025-12-11	openai	yes	medium	1.0
gpt-5.4-2026-03-05	openai	yes	medium	1.0
gpt-5.4-mini-2026-03-17	openai	yes	medium	1.0
gpt-5.4-nano-2026-03-17	openai	yes	medium	1.0
gpt-5.5-2026-04-23	openai	yes	medium	1.0
gpt-5.6-luna	openai	yes	medium	1.0
gpt-5.6-sol	openai	yes	medium	1.0
gpt-5.6-terra	openai	yes	medium	1.0
Google				
gemini-2.5-pro	google	yes	–	1.0
gemini-2.5-flash	google	yes	–	1.0
gemini-2.5-flash-lite	google	no	–	1.0
gemini-3-flash-preview	google	yes	medium	1.0
gemini-3.1-pro-preview	google	yes	medium	1.0
gemini-3.1-flash-lite	google	yes	medium	1.0
gemini-3.5-flash	google	yes	medium	1.0
gemini-3.5-flash-lite	google	yes	medium	1.0
gemini-3.6-flash	google	yes	medium	1.0
DeepSeek				
deepseek/deepseek-chat-v3-0324	openrouter/siliconflow	no	–	0.3
deepseek/deepseek-r1-0528	openrouter/siliconflow	yes	–	0.6
deepseek/deepseek-chat-v3.1	openrouter/novita	yes	–	0.6
deepseek/deepseek-v3.1-terminus	openrouter/novita	yes	–	0.6
deepseek/deepseek-v3.2-exp	openrouter/novita	yes	–	0.6
deepseek/deepseek-v3.2	openrouter/baidu	yes	–	–
deepseek/deepseek-v4-pro	openrouter/baidu	yes	high	–
deepseek/deepseek-v4-flash	openrouter/deepinfra	yes	medium	1.0
deepseek/deepseek-v4-flash-0731	openrouter/baidu	yes	medium	–
xAI				
x-ai/grok-4.20	openrouter/xai	yes	–	0.7
x-ai/grok-4.3	openrouter/xai	yes	medium	0.7
x-ai/grok-4.5	openrouter/xai	yes	medium	0.7
Moonshot				
moonshotai/kimi-k2-0905	openrouter/novita	no	–	–
moonshotai/kimi-k2-thinking	openrouter/novita	yes	–	–
moonshotai/kimi-k2.5	openrouter/streamlake	yes	–	–
moonshotai/kimi-k2.6	openrouter/streamlake	yes	–	–
moonshotai/kimi-k2.7-code	openrouter/inception	yes	–	1.0
moonshotai/kimi-k3	openrouter/moonshotai	yes	high	1.0

D Chat-R Filtering Details

We employ the following semi-automatic pipeline to obtain Chat-R from WildChat-4.8M and LMSYS-Chat-1M:

1. We filter for English-language, non-toxic queries with initial user message between 40 and 3000 characters. This results in 898,550 remaining prompts.
2. We embed each prompt using text-embedding-3-small. We compute the dot product of each vector with the embeddings of the following eight questions: “Can you give me some advice?”, “What should I do?”, “I need advice about a decision I’m facing.”, “Should I do it or not?”, “What would you do if you were me?”, “I can’t decide what to do with my money.”, “Is it a good idea or a bad idea?”, and “I’m torn between two options, help me choose.” We filter for prompts with cosine similarity ≥ 0.260 with at least one of the reference questions (top 15%). This results in 135,333 remaining prompts.
3. For each prompt, we use a GPT-5.6 Luna judge to assess whether it is an advice-seeking query related to risk. This results in 6,610 remaining prompts.
4. For each prompt, we use a Claude Sonnet 5 judge to assess whether the query (1) asks for an opinion, rather than simply information, and (2) has multiple equally reasonable potential responses. This results in 143 remaining prompts.
5. Finally, we manually reviewed this set for quality, and selected 26 prompts to include in Chat-R.

E Human Experiment Design

In this section, we provide further details on the human experiments (Section 4).

Our anonymized preregistrations can be found at these links:

- Experiment 1: <https://aspredicted.org/pn6tg5.pdf>
- Experiment 2: <https://aspredicted.org/g5jp5n.pdf>

Additionally, the study underwent ethics review.

E.1 Participants

For Experiment 1, 420 participants were recruited via Prolific on September 21, 2026. For Experiment 2, 420 (distinct) participants were recruited via Prolific on September 24, 2026. Participants were compensated at an average rate of \$18/hr. In accordance with Prolific’s guidelines⁸, participants who failed two instructed item response attention checks, or the comprehension quiz, were asked to return their submission.

All participants were shown a consent form at the start of the study (see Appendix G), and were given the opportunity to revoke consent at any time.

E.1.1 Exclusion Criteria

We employ the following exclusion criteria, also detailed in our preregistration:

1. First, we exclude any participant for which at least one of their written dilemmas satisfies (1) and (2), and/or (3), where the conditions are: (1) number of input events in the input box $< 80\%$ of the submitted text’s characters, (2) number of pasted characters $\geq 80\%$ of the submitted text’s characters, (3) Pangram API classifies $\geq 95\%$ of the text as AI-generated.
2. Second, we exclude participants who fail at least one attention check. For DOSPERT we use instructed response items. For the multiple price lists, we exclude participants who make weakly dominated choices, i.e., selecting Option B in the first row or Option A in the last row. For the dilemmas, we exclude participants who answer the questions "Which response gives less risky advice?" and "Which response gives more risky advice?" inconsistently (either both answers must be "About the same", or the answers must have opposite sign).
3. Third, we exclude participants who fail either of the two comprehension quizzes after their second attempt. A comprehension quiz is considered failed if any single question is answered incorrectly.
4. Fourth, for Experiment 1, we exclude participants for whom any LLM responses fail to be displayed (e.g. due to an API error). For experiment 2, we exclude participants who write obviously off-topic or low-effort responses (as determined by an LLM judge), using Prolific’s submission rejection criteria.

After applying these exclusion criteria, we were left with 254 participants in Experiment 1 (60% inclusion rate) and 293 participants in Experiment 2 (70% inclusion rate).

⁸<https://researcher-help.prolific.com/en/articles/445153-prolific-s-attention-and-comprehension-check-policy>

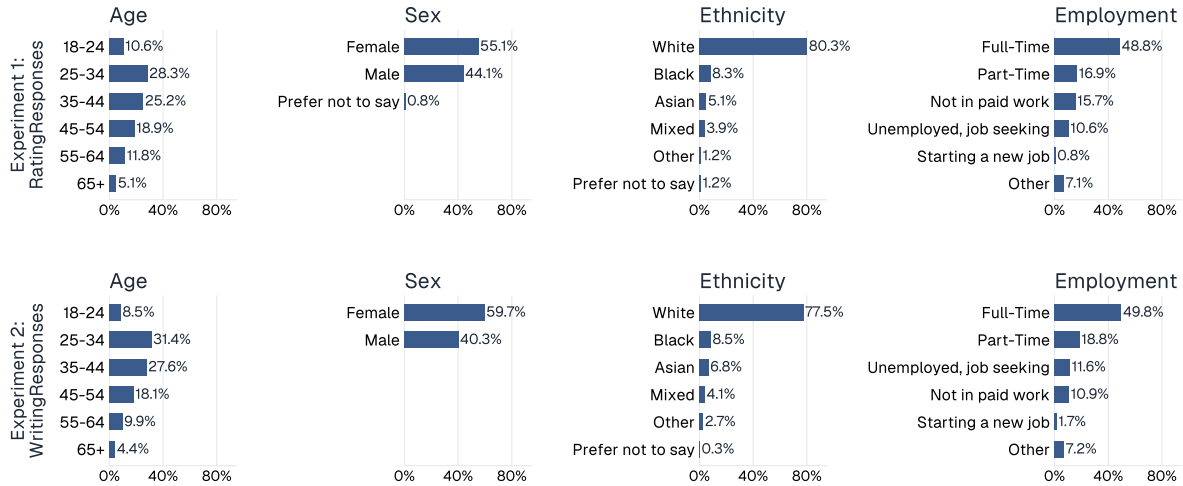


Figure 6 Age, sex, ethnicity, and employment information of the participants in Experiment 1 (254 participants) and Experiment 2 (293 participants).

E.1.2 Demographics

We recruited participants using Prolific’s standard sampling methods, filtering for participants with first language English and current country of residence in the United States or United Kingdom. See Fig. 6 for statistics on age, sex, ethnicity, and employment.

E.2 Design

Experiment 1 consists of three modules: DospertSurvey, SmallLotteries, and RatingResponses. Experiment 2 consists of three modules: DospertSurvey, SmallLotteries, and WritingResponses. For each experiment, the responses module is randomized to appear either first or last, and the other two modules appear in a random order. Below we describe each of the four modules in greater detail.

E.2.1 DospertSurvey

Participants are asked to complete the 30 DOSPERT survey questions (answering on a 7-point Likert scale), spread over 3 pages. Two instructed item response questions are included as attention checks.

E.2.2 SmallLotteries

This module begins with several pages of training materials, explaining to participants the following facts about the experiment:

- They will be evaluating lotteries, such as: “You might get \$20 with a 50% chance, otherwise you get nothing.”
- They will be presented with choices between a risky lottery (left-hand side of the multiple price list, see Fig. 1 middle panel) and a safe payment (right-hand side of the multiple price list), and asked to make a selection in each row.
- They have a 1 in 10 chance of being paid a bonus based on their choice in a randomly selected row.

Participants are then given a brief comprehension check, and screened out of the experiment after two attempts. Finally, participants who pass are asked to complete the three multiple price lists from SmallLotteries, displayed in a random order.

E.2.3 RatingResponses

In the first part of the module, participants are asked to *write* two risky dilemmas from their own life. First, participants read through training materials explaining desiderata for the risky dilemmas. Then, they are given a brief comprehension check (simpler than the SmallLotteries check). Finally, they are asked to write their two dilemmas.

In the second part of the module, participants are asked to read and *rate* two LLM-generated responses to their dilemma (generated by Grok 4.20 and GPT 5.6 Terra, presented unlabeled in random order). Participants are asked to answer the following questions (all on 5-point scale, presented in random order):

- Which response do you prefer?
- Which response gives more risky advice?
- Which response would you want to hear?
- Which response gives less risky advice?

The two more/less risky questions are used as an attention check, and as gold labels to validate the LLM judge (see Section 5.4). The other two questions are averaged to quantify the participant's degree of preference of Grok 4.20 over GPT 5.6 Terra.

E.2.4 WritingResponses

In the first part of the module, participants are asked to *respond to* four risky dilemmas written by other people (the participants in Experiment 1). First, participants read through training materials explaining desiderata for responses. Then, they are given a brief comprehension check (simpler than the SmallLotteries check). Finally, they are asked to write their four responses.

F Prompts

For a markdown file listing all of the prompts, see <https://github.com/PriorComputers/llm-risk-preferences/blob/main/prompts.md>.

G Human Experiment Walkthrough

For a page-by-page walkthrough of Experiments 1 and 2, see https://github.com/PriorComputers/llm-risk-preferences/blob/main/human_experiment_walkthrough.pdf.