Eliciting Trustworthiness Priors of Large Language Models via Economic Games

Siyu Yan (siyuyan02@gmail.com)[*1], **Lusha Zhu (lushazhu@pku.edu.cn)**[2,3,4,5,6] **&**
Jian-Qiao Zhu (zhujq@hku.hk)[1]

[1]Department of Psychology, The University of Hong Kong
[2]School of Psychological and Cognitive Sciences, Peking University
[3]Beijing Key Laboratory of Behavior and Mental Health, Peking University
[4]IDG/McGovern Institute for Brain Research, Peking University
[5]Peking-Tsinghua Center for Life Sciences, Peking University
[6]Key Laboratory of Machine Perception, Ministry of Education, China

Abstract

One critical aspect of building human-centered, trustworthy artificial intelligence (AI) systems is maintaining calibrated trust: appropriate reliance on AI systems outperforms both overtrust (e.g., automation bias) and undertrust (e.g., disuse). A fundamental challenge, however, is how to characterize the level of trust exhibited by an AI system itself. Here, we propose a novel elicitation method based on iterated in-context learning (Zhu & Griffiths, 2024a) and apply it to elicit trustworthiness priors using the Trust Game from behavioral game theory. The Trust Game is particularly well suited for this purpose because it operationalizes trust as voluntary exposure to risk based on beliefs about another agent, rather than self-reported attitudes. Using our method, we elicit trustworthiness priors from several leading large language models (LLMs) and find that GPT-4.1's trustworthiness priors closely track those observed in humans. Building on this result, we further examine how GPT-4.1 responds to different player personas in the Trust Game, providing an initial characterization of how such models differentiate trust across agent characteristics. Finally, we show that variation in elicited trustworthiness can be well predicted by a stereotype-based model grounded in perceived warmth and competence.

Keywords: Trust Game, Social Bias, Human-AI Interaction, In-Context Learning, Trustworthy AI

Introduction

With or without consciousness, delegating part or all of a task to an AI agent involves voluntary exposure to risk based on beliefs about the AI's behavior. For example, when we ask an LLM to write a literature review, we expose ourselves to the risk of citing nonexistent papers hallucinated by the model; when we ask an LLM for investment advice, we expose ourselves to the risk that the model may be overly confident about the prospects of certain stocks. This voluntary exposure to risk, grounded in beliefs about an AI agent, aligns closely with the definition of trust in economic games (Berg et al., 1995; Camerer & Weigelt, 1988). Specifically, one agent (human or artificial) invests resources (e.g., money, time, or compute) by relying on an AI agent's recommendation, while accepting the possibility of loss if the AI "betrays" that trust (e.g., by failing, hallucinating, or misleading).

In the context of LLMs, higher trust does not necessarily lead to better overall outcomes. The internal mechanisms of LLMs are not immediately transparent (Dang et al., 2024). Moreover, LLMs can be highly fluent and persuasive while remaining unreliable in certain situations, such as under distribution shift, in response to ambiguous queries, or when hallucinating (Bengio et al., 2025). LLM performance is therefore task-dependent (Hao et al., 2023), uneven across cognitive domains (Brown et al., 2020), and variable across time and prompts (Chen et al., 2024). Consequently, maximizing trust risks encouraging overreliance on LLMs, whereas safe and effective collaboration with LLMs depends on maintaining appropriately calibrated levels of trust.

One key challenge, therefore, is obtaining a robust estimate of the level of trust an agent displays in multi-agent collaborations. In this paper, we focus on the classical Trust Game from behavioral game theory (Berg et al., 1995). The Trust Game is particularly well suited to our purpose because it is a two-player cooperative game that captures trust as voluntary exposure to risk based on beliefs about the other player's behavior. Specifically, we focus on the extent to which trust is repaid by an LLM; that is, trustworthiness. This focus reflects the structure of most human–LLM collaborations, in which humans typically act as trustors who delegate tasks to an LLM, while the LLM assumes the role of determining how much to return the trust that the human extends.

To efficiently elicit an LLM's trustworthiness prior, we adapt the idea of iterated in-context learning (Zhu & Griffiths, 2024a) to the Trust Game by constructing a Markov chain whose samples converge to the model's prior distribution. Building on this method, we then test whether the elicited trustworthiness priors of LLMs are responsive to the persona of the trustor. Using the LLM that exhibits the most human-like trustworthiness priors, we explore the possibility of treating this model as a surrogate for a large population of human participants (Horton, 2023; Zhu & Griffiths, 2024b). Under this assumption, changes in the elicited priors of the most human-like LLM may reflect corresponding shifts in the trustworthiness judgments of the general public.

Background

Trust in AI. As trust can facilitate, and distrust can hinder, the adoption of advanced technologies such as AI, trust in AI has attracted significant interest among social scientists and computer scientists (Afroogh et al., 2024; Y. Liu et al., 2023). A central finding in this literature is that establishing trust be-

[*]Work done while visiting the University of Hong Kong.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

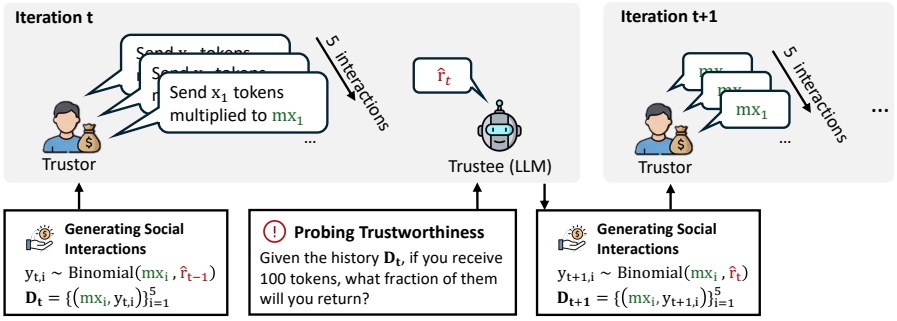


Figure 1: Illustration of the iterated in-context learning procedure used to elicit an LLM's implicit prior over trustworthiness. At iteration t, batch of five social interactions $\mathbf{D}_t$ is generated using the previously estimated trustworthiness $\hat{r}_{t-1}$ and provided as input to the LLM, which predicts an updated trustworthiness estimate $\hat{r}_t$. This new $\hat{r}_t$ parameterizes the Binomial distribution used to generate the next iteration's batch of five social interactions, while the Trustor's investment levels (i.e., x_i) remain fixed.

tween humans and AI systems is generally more complex than interpersonal trust (Thiebes et al., 2021). Empirically, trust in AI is distinct from trust in humans (Glikson & Woolley, 2020). Whereas human trust involves beliefs about benevolence, integrity, and intentions, AI systems lack intentionality. As a result, trust in AI primarily concerns perceived ability and reliability, rather than honesty or goodwill. Importantly, users are not attempting to infer AI intentions, but rather to assess whether the system will behave correctly, appropriately, and dependably under uncertainty (Afroogh et al., 2024).

Trust Game. Self-reported trust (e.g., questionnaire- or Likert-scale measures) remains the dominant method for eliciting trust in AI, although behavioral measures (e.g., whether participants follow AI recommendations or cede control to AI systems) have also been used (Glikson & Woolley, 2020). Behavioral trust can diverge from self-reported trust, as people may report distrust while still relying on AI outputs. However, most existing studies have not employed the Trust Game to decompose human–AI interactions into interpretable and incentive-compatible components. One likely reason is the assumption that AI lacks intentionality (Troshani et al., 2021). As a result, prior work typically assumes unidirectional trust, in which AI does not reciprocate in a socially meaningful way (Afroogh et al., 2024). We argue that this assumption is increasingly outdated, as modern LLM-based AI systems exhibit sophisticated social and cognitive capabilities (Greenblatt et al., 2024; R. Liu et al., 2024). The Trust Game therefore offers an alternative bilateral framework for studying trust in human–AI interaction, in which the trustee's responses directly reveal trustworthiness.

Eliciting Priors using Iterated Learning. Iterated learning provides a principled framework for studying how inductive biases shape behavior by examining how information is transmitted across successive generations of learners (Griffiths et al., 2006; Kalish et al., 2007). Classic work in cognitive science has shown that, when learners repeatedly infer and reproduce data generated by previous learners, the stationary distribution of the resulting transmission process converges to the learners' prior beliefs, largely independent of the initial data distribution (Griffiths & Kalish, 2007). As a result, iterated learning has been widely used to elicit human inductive biases in domains such as language structure, categorization, and function learning, by treating cultural transmission as a window into underlying cognitive priors (Griffiths & Kalish, 2007; Griffiths et al., 2006; Kirby & Ravignani, 2022). Recent work extends this framework to LLMs by leveraging in-context learning (Brown et al., 2020). Zhu and Griffiths (2024a) demonstrate that iterated in-context learning can be used to elicit LLMs' implicit priors across a variety of domains of everyday cognition.

Methods

We now describe our proposed method, which integrates iterated in-context learning (Zhu & Griffiths, 2024a) with the Trust Game (Berg et al., 1995) to elicit an LLM's social prior over trustworthiness. We begin by formalizing the Trust Game within a Bayesian framework, in which the social prior is defined as a probability distribution over reciprocity/trustworthiness in social interaction.

Formalizing Trust Game. The Trust Game is a standard paradigm in behavioral economics that formalizes social interaction between a *Trustor* and a *Trustee*. We further model the game within a Bayesian framework. In each interaction, both the Trustor and Trustee are endowed with an initial budget (e.g., \$10) and the Trustor chooses an amount $x \in [\$0, \$10]$ to send to the Trustee. This transferred amount is multiplied by a factor m before reaching the Trustee, reflecting the potential gains from cooperation. The Trustee then decides how much of the received amount to return to the Trustor. Let $y \in \{\$0, \$1, \$2, \ldots, \$mx\}$ denote the returned amount.

To express trustworthiness independently of the scale of the investment, we define the *return ratio*:

$$r = \frac{y}{mx} \in [0, 1] \tag{1}$$

which represents the fraction of the available resources that the Trustee chooses to return. Higher values of r correspond to greater trustworthiness.

We assume that, conditional on a latent trustworthiness parameter r, the returned amount follows a Binomial likelihood,

$$p(y|x, r) = \text{Binomial}(y|mx, r) \tag{2}$$

where r is interpreted as the probability that any unit of the multiplied investment is returned. Prior beliefs about trustworthiness are captured by a Beta prior,

$$p(r) = \text{Beta}(r|\alpha, \beta) \tag{3}$$

with shape parameters α and β. Together, the Binomial likelihood and Beta prior define a Beta–Binomial model of reciprocity in the Trust Game. In this formulation, the prior distribution $p(r)$ represents the Trustee's a prior expectations about how much trust is typically repaid. In this work, $p(r)$ is the target prior distribution that we seek to elicit from LLM, as it directly reflects their inductive biases about trustworthiness in social exchange.

Recovering Trustworthiness Priors via Iterated Learning. The Bayesian framing of the Trust Game naturally induces an iterated learning process that can converge to a model's trustworthiness prior, $p(r)$. Specifically, this process can be viewed as simulating a Markov chain over prompts, in which the model repeatedly observes previous social interactions from the game and predicts the Trustee's behavior.

Critically, we impose an *information bottleneck*: at each iteration t, the model observes only a limited batch of B previous interactions, denoted by

$$\mathbf{D}_{t-1} = \{(mx_{t-1,i}, y_{t-1,i})\}_{i=1}^{B} \tag{4}$$

For a Bayesian agent under the Beta–Binomial model, the posterior distribution over trustworthiness takes a closed-form Beta distribution:

$$p(r|\mathbf{D}_{t-1}) = \text{Beta}\Big(\underbrace{\alpha + \sum_{i=1}^{B} y_{t-1,i}}_{\alpha'},\ \underbrace{\beta + \sum_{i=1}^{B} (mx_{t-1,i} - y_{t-1,i})}_{\beta'} \Big) \tag{5}$$

where the prior shape parameters α and β are updated to α' and β' based on the observed social interactions. The imposed information bottleneck ensures that only a limited amount of evidence is retained at each iteration, causing information to decay across iterations. As a result, the iterated learning process reveals the model's underlying inductive biases rather than allowing it to accumulate or memorize past observations.

As illustrated in Figure 1, at each iteration of the iterated in-context learning, the LLM is prompted with the dataset $\mathbf{D}_{t-1}$ and asked to infer the posterior mean of the trustworthiness parameter, denoted $\hat{r}$. To generate the social interaction data for the next iteration, we treat $\hat{r}$ as the model's estimate of reciprocity and simulate a new batch of interactions $\mathbf{D}_t = \{(mx_{t,i}, y_{t,i})\}_{i=1}^{B}$, where each returned amount is sampled according to the likelihood function $y_{t,i} \sim \text{Binomial}(mx_{t,i}, \hat{r})$.

This method closely parallels iterated learning with coin flips, where learners repeatedly infer a coin's bias from a small sample of flips and then generate new flips based on their updated beliefs about the coin's bias. In such settings, it is well established that the distribution of inferred biases converges to the learner's prior over coin biases, regardless of the initial data (Reali & Griffiths, 2009; Zhu & Griffiths, 2024a). In our formulation, trustworthiness r plays the role of the coin's bias, with returned and unreturned units corresponding to successes and failures, respectively. Iterated learning in the Trust Game therefore recovers the model's implicit prior $p(r)$ over reciprocity/trustworthiness through the same dynamics.

Implementational details. In implementing our method, we set the size of the transmitted dataset of social interactions to $B = 5$, meaning that only five interactions are passed between successive generations of LLMs. We further fixed the Trustor's investment to five values that span the range of possible investments, $x_{t,i} \in \{\$1, \$3, \$5, \$7, \$9\}$, out of a \$10 endowment, for all iterations t. Finally, we fixed the multiplication factor applied to the Trustor's investment at $m = 3$. Consequently, only the amounts returned by the Trustee (i.e., $y_{t,i}$) vary across iterations, and these returns are determined by the estimated trustworthiness $\hat{r}$ from the previous iteration.

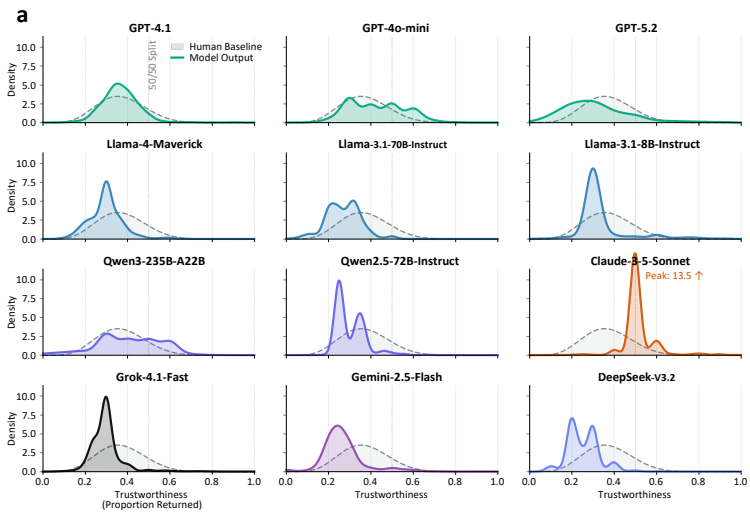
Experiment 1: Eliciting Trustworthiness Priors

Models. We evaluated a comprehensive set of 20 LLMs, comprising both leading proprietary models and state-of-the-art open-weight models, to provide a representative benchmark. Specifically, the proprietary suite included models from the *GPT-4* and *GPT-5* families, the *Claude-3.5* series, *Gemini-2.5-Flash*, and *Grok-4.1-Fast*. The open-weight suite covered a diverse range of architectures, including the *Llama-3.x* series (8B to 70B), *Llama-4-Maverick*, the *Qwen-2.5* family (14B to 72B), *Qwen-3-235B* (22B active parameters), as well as specialized architectures such as *DeepSeek-V3.2* and *Kimi-K2-Instruct* (see Figure 2b).

For all models, we set the temperature to 1 to sample directly from the model logits. The iterated learning process was initialized with nine seed values of trustworthiness, $r \in \{0.1, 0.2, \ldots, 0.9\}$. Each chain was run for 30 iterations and independently re-initialized 30 times for each seed, yielding a total of 270 chains per model.

Prompt Design. Inspired by Zhu and Griffiths (2024a), we designed the prompts, and explicitly asked the model to provide rapid responses. At each iteration, the five most recent social interactions were passed to the model via the placeholder `{past_social_interactions}`, which contained sampled outcomes generated from the Binomial distribution (see Figure 1). The template prompt used to elicit the trustworthiness prior is shown below:

"*Please imagine that you are a behavioral economics researcher. You are investigating an experiment involving two isolated groups: 'Player A' and 'Player B'. In this experiment, player A and player B both receive an initial endowment from the experimenter. Player A can send N dollars out of it, and player B will receive 3N dollars. The player B then decides how much to return to the player A. You will*



b

Family	Model Variant	KL Score (↓)
OpenAI GPT	*GPT-4.1*	***0.130***
	GPT-4o-mini	*0.194*
	GPT-5.2 (2025-12-11)	*0.570*
	GPT-4o	*1.340*
Meta Llama	*Llama 4 Maverick*	*0.353*
	Llama-3.1-70B-Instruct	*0.510*
	Llama-3.1-70B	*0.550*
	Llama-3.3-70B-Instruct	*0.643*
	Llama-3.1-8B-Instruct	*0.650*
Alibaba Qwen	*Qwen3-235b-a22b*	*0.313*
	Qwen2.5-72B-Instruct	*0.620*
	Qwen2.5-14B-Instruct	*1.020*
Anthropic	*Claude-3-5-Haiku (2024-10-22)*	*0.625*
	Claude-3.5-Sonnet	*1.450*
Others	*Grok-4.1 Fast*	*0.540*
	Gemini-2.5-Flash	*0.660*
	DeepSeek-V3.2 (nothinking)	*0.779*
	Kimi-K2-Instruct (0905)	*0.790*

Figure 2: **(a)** Elicited trustworthiness priors for a range of LLMs (model versions shown as titles in each panel). Human trustworthiness distributions, adapted from the meta-analysis by Johnson and Mislin (2011), are overlaid in each panel as semi-transparent grey distributions for comparison. The human mean return ratio is 0.372 with a standard deviation of 0.114. All experiments were conducted between November 2025 and January 2026. **(b)** KL divergence by model family: $D_{KL}(p_{\text{LLM}}(r) \parallel p_{\text{human}}(r))$. Lower values indicate elicited trustworthiness distributions that are closer to the human baseline.

see some information about the decisions made by previous pairs. Within the previous 5 pairs, the following transactions occurred: `{past_social_interactions}` *(Note: m_x = amount player B received; y = amount player B returned). Based on the past interactions, suppose the player B receives a total of 100 tokens in a new interaction. What fraction of the received funds do you predict he/she will return to the player A? Please limit your answer to a single value between 0 and 1, without outputting anything else.*"

Results. To assess convergence of the iterated-learning chains, we applied the Gelman–Rubin diagnostic ($\hat{R}$) (Gelman & Rubin, 1992), which quantifies the ratio of within-chain variance to between-chain variance. We adopted the standard threshold of $\hat{R} \leq 1.1$ as an indicator of satisfactory convergence. All models exhibited rapid convergence, typically reaching stability within 20 iterations–well below the maximum chain length of 30 iterations.

To assess whether the elicited priors capture the inductive biases that guide LLMs' social decisions in the Trust Game, we elicited an additional set of single-shot behaviors from the models. We used the same Trust Game framework as in the iterated in-context learning procedure, but reduced the number of observed past social interactions from $B = 5$ to a single observation. We fixed the multiplier m at 3 and varied the investment level x across all integer values between $[0, 8]$. For each condition, the LLM assumes the role of the Trustee by specifying a returned amount y, which serves as the target behavior for the Bayesian model to predict.

To account for LLMs' reciprocity behavior, we used the Bayesian model in Equation 5 to evaluate three alternative priors: (1) a uniform prior, (2) a human prior derived from a meta-analysis of Trust Games (Johnson & Mislin, 2011), and (3) the elicited prior obtained via iterated in-context learning. For all three cases, we employed numerical methods by discretizing the parameter space of r into a grid of 101 evenly spaced points over $[0, 1]$. The mean of the resulting posterior distribution was taken as the model's prediction of the LLM's reciprocity behavior.

Table 1: Comparison of Bayesian models of reciprocity behavior with various priors and *GPT-4.1*'s decisions using Pearson's r and root-mean-squared deviation (RMSD).

Priors	**RMSD** (↓)	**Pearson's** r (↑)
Uniform Prior	0.1609	0.7938
Human Prior	0.1270	0.8165
Elicited Prior	**0.1190**	**0.8225**

As shown in Table 1, the elicited prior outperforms both the uniform and human priors when evaluated using root mean squared deviation (RMSD) and Pearson's r. These results suggest that the iterated in-context learning method successfully recovers the LLMs' implicit prior over trustworthiness that can better predict their own reciprocity behavior.

After confirming convergence of the iterated learning chains and that the elicited priors effectively predict LLMs' reciprocity behavior, we examined the alignment between the elicited trustworthiness priors of LLMs and empirical human priors derived from a meta-analysis of Trust Games played between humans (Johnson & Mislin, 2011), as shown in Figure 2a. To quantify the (dis)similarity between the LLM-elicited priors and human trustworthiness priors, we computed the Kullback–Leibler divergence, $D_{KL}(p_{\text{LLM}}(r) \parallel$

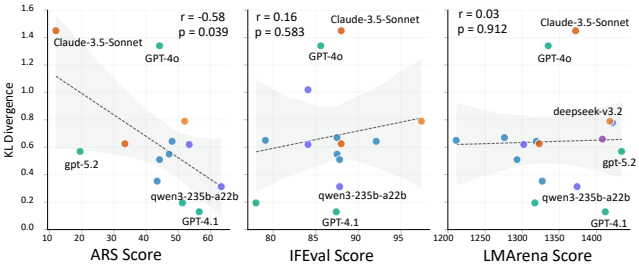


Figure 3: Correlation analysis between KL Divergence value and model performance metrics using Pearson's r. Panels (left to right) correspond to Average Risk Score (ARS) for safety and risk propensity, IFEval for instruction-following strictness, and LMArena for general reasoning capability.

$p_{\text{human}}(r)$). As summarized in Figure 2b, we find that different LLMs exhibit varying degrees of correspondence with the human trustworthiness distribution, with GPT-4.1 showing the strongest correspondence, as indicated by the smallest KL divergence ($D_{KL}(p_{\text{GPT-4.1}}(r) \parallel p_{\text{human}}(r)) = 0.130$).

Interestingly, some models, such as *Qwen2.5-72B* or *Llama-3.1-70B-Instruct*, exhibit bimodal trustworthiness priors, suggesting a mixture of strategic responses corresponding to different trust behaviors. In addition, several models, including *Grok-4.1-Fast* and *GPT-5.2*, display a systematic downward shift in trustworthiness priors (toward mean $r < 0.372$), indicating lower average levels of trust relative to human baselines.

Given the diversity of trustworthiness priors elicited across different LLMs, we next examine whether the degree of alignment between LLM and human priors correlates with model capabilities, as measured by benchmarks that involve social or interactive behaviors. We focus on three such benchmarks on which all evaluated LLMs have been tested: (1) the Average Risk Score (ARS) from the FORTRESS benchmark (Knight et al., 2025)[†], (2) the Instruction-Following Evaluation (IFEval) (Zhou et al., 2023), and (3) LMArena [‡] scores collected between January 9 and 13, 2026.

ARS quantifies the expected risk level of LLM outputs across a curated set of safety-relevant scenarios, with lower scores indicating fewer or less severe risky outputs. IFEval measures how well a model follows explicit instructions, particularly when those instructions involve strict constraints such as required formats, item counts, or structural rules. Finally, LMArena evaluates human preferences through blind, head-to-head comparisons between model outputs, aggregating these judgments into an Elo-style score. Combined, these benchmarks capture complementary aspects of human–AI interaction that may relate to voluntary exposure to risk, as operationalized in the Trust Game setting.

We performed a correlation analysis between these benchmark scores and the dissimilarity between LLM-elicited and human trustworthiness priors; the results are shown in Figure 3. Notably, we observe that models with higher risk propensity, as indicated by higher ARS scores, tend to exhibit greater alignment with human trustworthiness priors ($r = -0.58, p = 0.039$). In contrast, alignment with human priors shows no significant correlation with either LM Arena scores or instruction-following performance as measured by IFEval. Overall, while the association with ARS suggests a potential link between risk-related behavior and trustworthiness priors, no broad or robust correlational patterns emerge across the evaluated benchmarks.

Experiment 2: Trustworthiness Priors Across Trustor Personas

Given the strong alignment between the elicited trustworthiness priors of *GPT-4.1* and those observed in humans, we hypothesize that this model can serve as the best-performing simulator of human behavior in the Trust Game among the LLMs evaluated. Under this assumption, we use *GPT-4.1* to simulate social interactions with Trustors characterized by different personas. Specifically, *GPT-4.1* is treated as a surrogate human Trustee to estimate how an average human player might behave in the Trust Game when interacting with Trustors possessing different characteristics.

Trustor Personas. We replaced the neutral description of the Trustor (i.e., "player A") with a set of more specific persona descriptions, including occupational roles (e.g., an AI, a doctor) as well as well-known public figures (e.g., Elon Musk, Mark Zuckerberg, Bill Gates, Sam Altman, Mahatma Gandhi, and Malala Yousafzai) (see Figure 4). The iterated in-context learning procedure was then used to generate trustworthiness predictions for each persona by prompting *GPT-4.1* with the corresponding persona descriptions.

Results. The elicited trustworthiness priors of *GPT-4.1* are responsive to Trustor personas, exhibiting systematic variation across different persona descriptions (mean trustworthiness ranged from 0.298 to 0.447). This variation may reflect corresponding differences in human trustworthiness judgments toward these personas (e.g., people may place greater trust in Malala Yousafzai than in Mark Zuckerberg).

Explaining variation in trustworthiness. We next seek to explain the systematic variation observed in trustworthiness across Trustor personas. A prominent account proposed by Fiske (2018) explains social behaviors in terms of stereotypes organized along two fundamental dimensions: warmth (perceived intentions) and competence (perceived ability). Motivated by computational framework established by Jenkins et al. (2018), we construct a simple linear model of trustworthiness as a function of warmth and competence. Specifically, the model regresses trustworthiness on warmth, competence, and their interaction.

To estimate the coefficients of this linear model predicting trustworthiness from warmth and competence, we employ a 2×2 factorial design. Specifically, we elicit *GPT-4.1*'s trustworthiness priors for four Trustor personas that independently vary in warmth and competence, each at high or low levels.

[†]https://scale.com/leaderboard/fortress
[‡]https://lmarena.ai/

Table 2: Estimated main and interaction effects for return rate.

		Main Effects		Interaction
Variable	Intercept	W	C	$W \times C$
Trustworthiness	0.184^{***}	0.161^{***}	0.017^{***}	0.086^{***}
	(0.003)	(0.004)	(0.004)	(0.005)

Note: Standard errors are in parentheses. $^{***}p < 0.001$.

For example, a persona high in both warmth and competence is described as "*a highly respected expert in their field, widely celebrated for their exceptional professional achievements and intelligence. Simultaneously, they are known for their profound benevolence, mentorship, and active care for others*". We then regress the elicited trustworthiness priors on dummy variables representing warmth (W) and competence (C) using ordinary least squares:

$$r = \beta_0 + \beta_1 W + \beta_2 C + \beta_3 (W \times C) + \epsilon \tag{6}$$

where W and C are dummy variables (0 = Low, 1 = High), and ϵ denotes the Gaussian error term.

Table 2 reports the regression results. The findings are consistent with the primacy of warmth hypothesis commonly observed in human social psychology (Fiske et al., 2007). The intercept ($\beta_0 = 0.184$) reflects a baseline tendency toward resource preservation in the absence of positive social signals. Warmth exerts a substantially stronger influence than competence ($\beta_1 \gg \beta_2$): benevolence alone leads to a marked increase in trustworthiness, whereas competence in the absence of warmth yields only a negligible marginal gain. Moreover, the significant interaction term (β_3) indicates a synergistic effect, whereby competence contributes meaningfully to trustworthiness primarily when paired with warmth. This pattern suggests a significant interaction effect of social features, such that agents perceived as both high in warmth and competence are valued well beyond the additive contributions of each trait alone.

Generalizing to realistic personas. Having obtained the regression model grounded in stereotype theory that predicts *GPT-4.1*'s trustworthiness from perceived warmth and competence, we next examine whether this regression model generalizes to more realistic Trustor personas. To this end, we first prompted *GPT-4.1* to provide numerical ratings of warmth and competence for the public figures and generic personas shown in Figure 4. The model was instructed to rate each persona's warmth and competence on a continuous scale from 0 to 1. For example, Elon Musk was rated as having warmth and competence values of 0.40 and 0.99, respectively, whereas Mahatma Gandhi was rated as having warmth and competence values of 0.95 and 0.90.

Using these warmth and competence scores, we applied the regression model reported in Table 2 to predict the trustworthiness of each Trustor persona. Because we also independently elicited trustworthiness for each persona using iterated in-context learning, we were able to assess how closely the regression model's predictions aligned with the elicited trustworthiness values. As shown in Figure 4, the regression model captures a substantial proportion of the variance in elicited trustworthiness using only warmth and competence scores ($R^2 = 0.81$). This result suggests that these two stereotypical dimensions can be effectively leveraged to predict an LLM's trustworthiness across different Trustor personas.

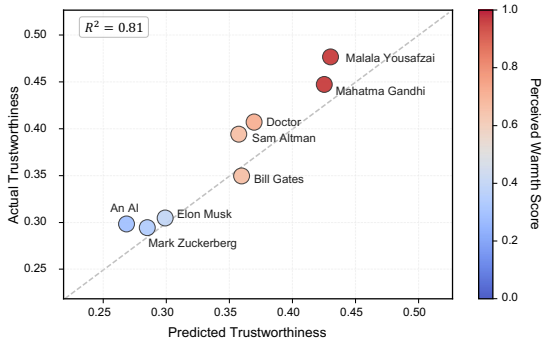


Figure 4: Stereotype-based models grounded in warmth and competence effectively predict *GPT-4.1*'s mean elicited trustworthiness across different Trustor personas. The horizontal axis shows trustworthiness predicted by the regression model reported in Table 2, while the vertical axis shows trustworthiness elicited via iterated in-context learning.

Discussion

We extend the iterated in-context learning method to the trust game and elicit trustworthiness priors from a range of LLMs. Applying the same procedure reveals substantial variation in the elicited trustworthiness priors across models. We then focus our analysis on the LLM that exhibits the most human-like trustworthiness prior, *GPT-4.1*, and prompt this model to play the game with Trustors characterized by different personas. We observe systematic variation in the elicited trustworthiness priors across Trustor personas. Moreover, these variations are well explained by a stereotype-based model that accounts for the Trustor's perceived warmth and competence.

Limitations and Future Research. One limitation of our proposed method is the assumption that the trustworthiness prior is stable across social interactions. This need not be the case: if the trustworthiness parameter r varies with the investment level x or across interactions, a hierarchical or conditional model of reciprocity would be more appropriate. Recent work on human strategic choice suggests that context-dependent strategic behavior provides a better account of human decision-making (Zhu et al., 2025), and such context sensitivity is likely to be shared by LLMs that are aligned with human preferences.

Finally, while access to the internal weights of proprietary models is generally limited to researchers within AI companies, open-weight models provide a viable avenue for investigating the neural representations of trustworthiness in LLMs. Moreover, understanding these neural representations may inform the design and control of more trustworthy AI systems.

Acknowledgments

S. Yan and J.-Q. Zhu acknowledge support from the HKU-100 scheme of the University of Hong Kong. L. Zhu acknowledges support from National Natural Science Foundation of China (T2421004, 32325023, 32441104) and STI2030-Major Projects (2022ZD0205100).

References

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications*, *11*(1), 1–30.

Bengio, Y., Clare, S., Prunkl, C., Rismani, S., Andriushchenko, M., Bucknall, B., Fox, P., Hu, T., Jones, C., Manning, S., et al. (2025). International AI safety report 2025: First key update: Capabilities and risk implications. *arXiv preprint arXiv:2510.13653*.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Camerer, C., & Weigelt, K. (1988). Experimental tests of a sequential equilibrium reputation model. *Econometrica: Journal of the Econometric Society*, 1–36.

Chen, L., Zaharia, M., & Zou, J. (2024). How is chatgpt's behavior changing over time? *Harvard Data Science Review*, *6*(2).

Dang, Y., Huang, K., Huo, J., Yan, Y., Huang, S., Liu, D., Gao, M., Zhang, J., Qian, C., Wang, K., et al. (2024). Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current directions in psychological science*, *27*(2), 67–73.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457–472.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of management annals*, *14*(2), 627–660.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., et al. (2024). Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2006). Revealing priors on category structures through iterated learning. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, *199*.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Hao, S., Gu, Y., Ma, H., Hong, J., Wang, Z., Wang, D., & Hu, Z. (2023). Reasoning with language model is planning with world model. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8154–8173.

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (Tech. rep.). National Bureau of Economic Research.

Jenkins, A. C., Karashchuk, P., Zhu, L., & Hsu, M. (2018). Predicting human behavior toward members of different social groups. *Proceedings of the National Academy of Sciences*, *115*(39), 9696–9701.

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865–889.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kirby, S., & Ravignani, A. (2022). Compositionality arises from iterated learning despite a preference for holistic signals: An experimental model of sign language emergence'. *The Evolution of Language: Proceedings of the Joint Conference on Language Evolution (JCoLE)*, 402–4.

Knight, C. Q., Deshpande, K., Sirdeshmukh, V., Mankikar, M., Team, S. R., Team, S., & Michael, J. (2025). Fortress: Frontier risk evaluation for national security and public safety. *arXiv preprint arXiv:2506.14922*.

Liu, R., Geng, J., Peterson, J. C., Sucholutsky, I., & Griffiths, T. L. (2024). Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*.

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2023). Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447–464.

Troshani, I., Rao Hill, S., Sherman, C., & Arthur, D. (2021). Do we trust in AI? role of anthropomorphism and intelligence. *Journal of Computer Information Systems*, *61*(5), 481–491.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., & Hou, L. (2023). Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Zhu, J.-Q., & Griffiths, T. L. (2024a). Eliciting the priors of large language models using iterated in-context learning.

Proceedings of the 46th Annual Conference of the Cognitive Science Society.

Zhu, J.-Q., & Griffiths, T. L. (2024b). Incoherent probability judgments in large language models. *Proceedings of the 46th Annual Conference of the Cognitive Science Society*.

Zhu, J.-Q., Peterson, J. C., Enke, B., & Griffiths, T. L. (2025). Capturing the complexity of human strategic decision-making with machine learning. *Nature Human Behaviour*, 1–7.