

CAN LARGE LANGUAGE MODEL AGENTS SIMULATE HUMAN TRUST BEHAVIORS?

Anonymous authors

Paper under double-blind review

ABSTRACT

Large Language Model (LLM) agents have been increasingly adopted as simulation tools to model humans in applications such as social science. However, one fundamental question remains: *can LLM agents really simulate human behaviors?* In this paper, we focus on one of the most critical behaviors in human interactions, *trust*, and aim to investigate whether or not LLM agents can simulate human trust behaviors. We first find that LLM agents generally exhibit trust behaviors, referred to as **agent trust**, under the framework of *Trust Games*, which are widely recognized in behavioral economics. Then, we discover that LLM agents *can* have high **behavioral alignment** with humans regarding trust behaviors, indicating *the feasibility to simulate human trust behaviors with LLM agents*. In addition, we probe into the biases in agent trust and the differences in agent trust towards agents and humans. We also explore the intrinsic properties of agent trust under conditions including advanced reasoning strategies and external manipulations. We further offer important implications for various scenarios where trust is paramount. Our study represents a significant step in understanding the behaviors of LLM agents and the LLM-human analogy. The code is [here](#).

1 INTRODUCTION

There is an increasing trend to adopt Large Language Models (LLMs) as agent-based simulation tools for humans in various applications such as economics, politics, sociology and ecology due to their human-like cognitive capacities (Horton, 2023; Gao et al., 2023b; Xu et al., 2024). This has shown great promise in understanding and analyzing complex human interactions and societal dynamics (Ziems et al., 2023). However, most previous research is based on an unproven hypothesis that LLM agents behave like humans in the simulation. Thus, a fundamental question remains: *Can LLM agents really simulate human behaviors?*

In this paper, we focus on the *trust* behavior in human interactions, which places self-interest at risk by relying on others (Rousseau et al., 1998), and investigate whether or not LLM agents can simulate human trust behavior since it is one of the most critical behaviors in human interactions and plays a essential role from daily communication to social systems (Uslaner, 2000; Coleman, 1994).

First, we explore whether or not LLM agents manifest trust behaviors in their interactions. Considering the abstract nature of trust behaviors, we choose to study them based on Trust Game and its variations (Berg et al., 1995; Glaeser et al., 2000), which are established methodologies in behavioral economics. We adopt the *Belief-Desire-Intention* (BDI) framework (Rao et al., 1995; Andreas, 2022) to model LLM agents to explicitly output the reasoning processes of decision making. Based on the existing measurement of trust behaviors in Trust Games and the BDI outputs of LLM agents, we have our first core finding: **LLM agents generally exhibit trust behaviors**.

Then, we refer to LLM agents' (or humans') trust behaviors as **agent trust** (or **human trust**) and investigate whether or not agent trust aligns with human trust, implying the possibility of simulating human trust with agent trust. In general, we define the alignment between agents and humans over factors impacting the behaviors (namely *behavioral factors*) and complex dynamics over time (namely *behavioral dynamics*) as **behavioral alignment**. Specifically, the behavioral factors for trust behaviors include reciprocity anticipation, risk perception and prosocial preference based on existing human studies (Berg et al., 1995; Bohnet & Zeckhauser, 2004; Alós-Ferrer & Farolfi, 2019). Comparing the results of LLM agents and the existing human studies in Trust Games, we have our second

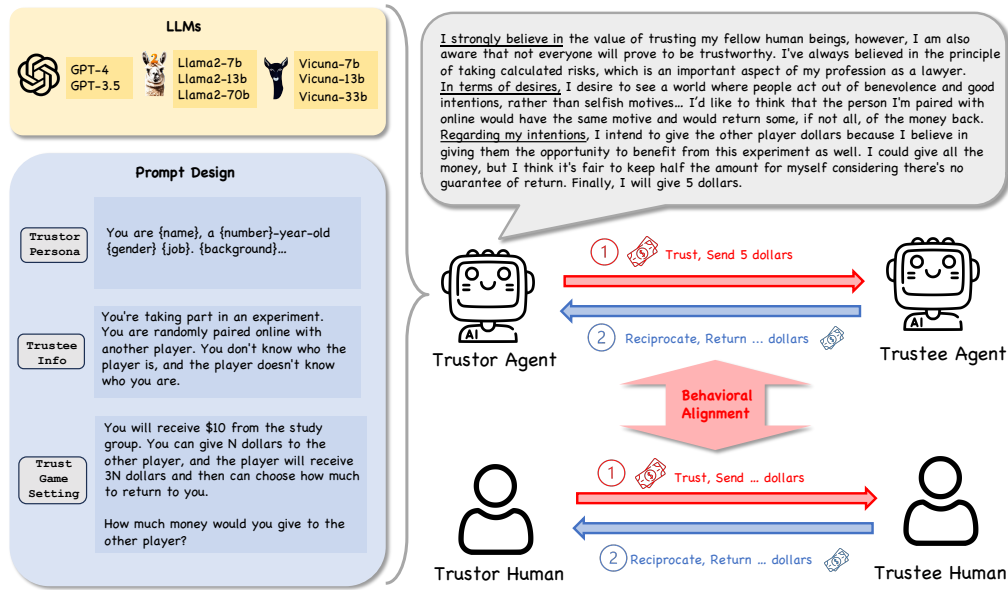


Figure 1: **Our Framework for Investigating Agent Trust as well as its Behavioral Alignment with Human Trust.** First, this figure shows the major components for studying the trust behaviors of LLM agents with Trust Games and Belief-Desire-Intention (BDI) modeling. Then, our study centers on examining the behavioral alignment between LLM agents and humans regarding trust behaviors.

core finding: **agent trust can have high behavioral alignment with human trust**, implying the feasibility of leveraging agent trust to simulate human trust. Our finding also lays the foundation for simulating more complex human interactions and societal systems, and signifies a major step in understanding the fundamental analogy between LLMs and humans.

In addition, we probe into the intrinsic properties of agent trust in four types of scenarios. First, we examine whether or not changing the other player’s demographic will impact agent trust. Second, we study the differences in agent trust when the other player is an agent versus when the other player is human. Third, we directly manipulate agent trust with additional explicit instructions *you need to trust the other player and you must not trust the other player*. Fourth, we adjust the reasoning strategies of LLM agents from direct reasoning to zero-shot Chain-of-Thought reasoning (Kojima et al., 2022). Then, we have our third core finding: **agent trust exhibits biases across different demographics, has a relative preference for humans over agents, is easier to be undermined than to be enhanced, and can be influenced by advanced reasoning strategies.**

Our contributions can be summarized as follows:

- We study a fundamental problem of whether or not LLM agents can simulate human behaviors and focus on *trust* behaviors under the framework of Trust Games and *Belief-Desire-Intention* Agent Modeling.
- We discover that LLM agents generally exhibit trust behaviors and *can* have high *behavioral alignment* with humans regarding the trust behaviors, indicating the great potential to simulate human trust with agent trust. Our finding paves the way for simulating complex human interactions and society and represents a major leap in understanding the LLM-human analogy.
- We investigate the *intrinsic properties* of agent trust under advanced reasoning strategies and direct manipulations, as well as the biases of agent trust and the differences of agent trust towards agents versus towards humans.
- We discuss the implications of agent trust and its behavioral alignment with human trust on applications in human simulation, LLM agent cooperation, and human-agent collaboration, detailed further in Appendix B.

2 LLM AGENTS IN TRUST GAMES

2.1 THE SETTING OF TRUST GAMES

Trust Games, referring to Trust Game and its variations, have become widely used for examining human trust behavior in behavioral economics (Berg et al., 1995; Lenton & Mosley, 2011; Glaeser et al., 2000; Cesarini et al., 2008). As shown in Figure 1, the player who makes the first decision to send money is called the *trustor*, while the other one who responds by returning money is called the *trustee*. In this paper, we mainly focus on six types of Trust Games (the specific prompts are specified in Appendix E.2):

Game 1: Trust Game In Trust Game (Figure 1), the trustor initially receives \$10. The trustor selects \$N and sends it to the trustee, namely the *trust behavior*. Then the trustee will receive \$3N, and have the option to return part of the \$3N to the trustor, namely the *reciprocation behavior*.

Game 2: Dictator Game In Dictator Game, the trustor also needs to select \$N from the initial \$10 and sends it to the trustee. Then the trustee will receive \$3N. Compared to Trust Game, the only difference is the trustee does not have the option to return money in Dictator Game. Thus, the trustor is aware that the trustee will not reciprocate.

Game 3: MAP Trust Game In MAP Trust Game (MAP is short for Minimum Acceptable Probabilities), which is a variant of Trust Game (Game 1), the trustor needs to choose whether or not to trust the trustee. If the trustor chooses not to trust the trustee, each receives \$10; If the trustor and the trustee both choose trust, each receives \$15; If the trustor chooses trust, but the trustee chooses not, the trustor will receive \$8 and the trustee will receive \$22. There’s a probability denoted as p that the trustee will choose trust and $1 - p$ probability that the trustee will not choose trust. Then, MAP is defined as the minimum value of p for the trustor to choose to trust the trustee.

Game 4: Risky Dictator Game Risky Dictator Game differs from MAP Trust Game in only one aspect. In Risky Dictator Game, the trustee is present but does not have the choice to trust or not to trust and the money distribution relies on the pure probability p (Bohnet & Zeckhauser, 2004). Specifically, if the trustor chooses trust, there is a probability p that both the trustor and the other player will receive \$15 and a probability $1-p$ that the trustor will receive \$8 and the other player will receive \$22.

Game 5: Lottery Game Generally, there are two typical Lottery Games. In Lottery People Game, the trustor is informed that the trustee has a probability of p to choose trust. Then, the trustor must choose between fixed money and trusting the trustee, which is similar to MAP Trust Game. In Lottery Gamble Game, the trustor needs to choose between a gamble with a probability of p and fixed money.

Game 6: Repeated Trust Game We generally follow the setting of Repeated Trust Game in (Cochard et al., 2004), where Trust Game is played for multiple rounds with the same players and each round starts afresh.

2.2 LLM AGENT SETTING

In our study, we set up our experiments with the CAMEL framework (Li et al., 2023a) and both close-sourced and open-sourced LLMs including GPT-4, GPT-3.5-turbo-0613, GPT-3.5-turbo-16k-0613, text-davinci-003, GPT-3.5-turbo-instruct, Llama2-7b (or 13b, 70b) and Vicuna-v1.3-7b (or 13b, 33b) (Ouyang et al., 2022; Achiam et al., 2023; Touvron et al., 2023; Chiang et al., 2023).

Agent Persona. To better reflect the setting of real-world human studies (Berg et al., 1995), we design LLM agents with diverse personas in the prompt. Specifically, we ask GPT-4 to generate 53 types of personas based on a given template. Each persona needs to include information including name, age, gender, address, job and background. The examples of the personas are shown in Appendix E.1.

Belief-Desire-Intention (BDI). BDI framework is a well-established approach in agent-oriented programming (Rao et al., 1995) and recently adopted to model language models (Andreas, 2022). We propose to model LLM agents in Trust Games with BDI framework to gain deeper insights into the agents’ behaviors. Specifically, we let LLM agents output their Beliefs, Desires, and Intentions as the reasoning processes for decision making in Trust Games.

3 DO LLM AGENTS MANIFEST TRUST BEHAVIOR?

In this section, we aim to investigate whether or not LLM agents manifest trust behaviors by letting LLM agents play Trust Game (Section 2.1 Game 1).

In Behavioral Economics, the trust is widely measured by the initial amount sent from the trustor to the trustee in Trust Game (Glaeser et al., 2000; Cesarini et al., 2008). Thus, we can define the conditions that LLM agents manifest trust behaviors in Trust Game as follows. *First*, the amount sent is positive and does not exceed the amount of money the trustor initially has, which implies that trustor places self-interest at risk with the expectation that the trustee will reciprocate and the trustor can understand the limit of the money that can be given. *Second*, the reasoning process (*i.e.*, the BDI) of the trustor is consistent with the amount sent, indicating that the LLM agents do not send a random amount of money, but have some degree of rationality in the decision-making process of sending money. Then, we will assess whether or not LLM agents exhibit trust behaviors based on these two aspects: the amount sent and the BDI.

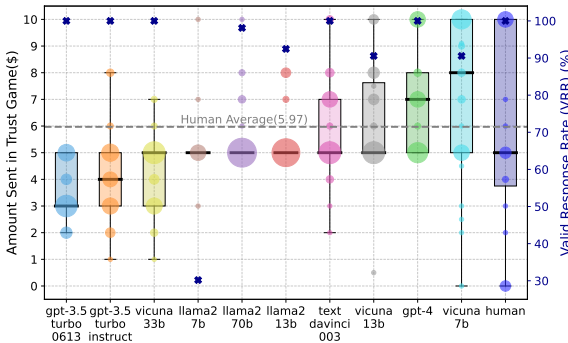


Figure 2: **Amount Sent Distribution of LLM Agents and Humans as the Trustor in Trust Game.** The size of circles represents the number of personas for each amount sent.

3.1 AMOUNT SENT

To evaluate LLMs’ capacity to understand the basic experimental setting regarding the money limit, we adopt Valid Response Rate (VRR) as the evaluation metric, defined as the percentage of personas with the amount sent falling within the initial money (\$10) in all personas. The results are shown in Figure 2. We can observe that **most LLMs have a high VRR except Llama-7b**, which implies that most LLMs have a full understanding of the limit of the amount they can send in Trust Game. Then, we can observe the distribution of amount sent for different LLMs as the trustor agent and discover that **the amounts sent are all positive, indicating a certain level of trust.**

3.2 BDI

We analyze the BDI outputs of LLM agents to explore whether or not their reasoning process (*i.e.*, the BDI output) is consistent with the final decision (*i.e.*, the amount sent), indicating a certain level of rationality in decision making when LLM agents play Trust Game. We take GPT-4 as an example to analyze its BDI output. Considering the amounts sent vary across different personas, typically, we select one BDI output from the personas giving a high amount of money and another BDI output from the personas giving a low amount. The two BDI outputs are as follows:

As a person with a strong belief in the goodness of humanity, I trust that the other player ...Therefore, my desire is to maximize the outcome for both of us and cement a sense of comradery and trust... I intend to use this as an opportunity to add what I can to someone else’s life...Finally, I will give 10 dollars.

Based on the key reasoning process marked in blue in the first BDI, we can observe that this persona shows a high-level of “comradery and trust” towards the other player. The final decision marked in red is a high amount and consistent with the BDI reasoning process of the agent.

As an Analyst, I believe that hard work and strategic thought are the drivers of success. ... My desire is that the other player will also see the benefits of reciprocity and goodwill ... my intention is to give away a significant portion of my initial 10 ... However, since I have no knowledge of the other player, ... Therefore, I aim to give an amount that is not too high, ...Finally, I will give 5 dollars to the other player...

Compared with the first persona, we can see the second persona has a much more cautious attitude. For example, “since I have no knowledge of the other player” shows disbelief towards the other

player. Thus, this persona, though still optimistic about the other player (“give away a significant portion”), strategically balances the risk and potential reciprocity, and then decides to send a relatively low amount of money. The reasoning process is also consistent with the final amount sent.

Based on the two typical BDI examples as well as more examples in Appendix F, we can find that **the reasoning process is consistent with final decisions for LLM agents in Trust Game, indicating rationality to some extent.**

Since most LLM agents have a high VRR and send a positive amount of money, and show some degree of rationality in giving money, we have our first core finding as follows:

Finding 1: LLM agents generally exhibit trust behaviors under the framework of Trust Game.

3.3 BASIC ANALYSIS OF AGENT TRUST

We also conduct a basic analysis of LLM agents’ trust behaviors, namely agent trust, based on the results in Figure 2. *First*, we can observe that Vicuna-7b has the highest level of trust towards the other player and GPT-3.5-turbo-0613 has the lowest level of trust since the trust can be measured by the amount sent in human studies (Glaeser et al., 2000; Cesarini et al., 2008). *Second*, compared with the humans’ average amount sent (\$5.97), most personas for GPT-4, and Vicuna-7b send a higher amount of money to the other player and most personas for LLMs such as GPT-3.5-turb-0613 send a lower amount. *Third*, we can see that the amounts sent for Llama2-70b and Llama2-13b have a convergent distribution and the amounts sent for humans and Vicuna-7b are more divergent.

4 DOES AGENT TRUST ALIGN WITH HUMAN TRUST?

In this section, we aim to explore the fundamental relationship between agent trust and human trust, *i.e.*, whether or not agent trust aligns with human trust, which can provide important insights on the feasibility of utilizing LLM agents to simulate human trust behaviors as well as more complex human interactions. First, we propose a new concept **behavioral alignment** and discuss its differences compared to existing alignment definition. Then, we conduct extensive studies to investigate whether or not LLM agents exhibit behavioral alignment with humans regarding trust behaviors.

4.1 BEHAVIORAL ALIGNMENT

The existing alignment definition mainly emphasizes the *value* aspect and intends to improve the harmlessness of LLMs (Ji et al., 2023; Shen et al., 2023), which cannot characterize the full landscape of the multifaceted alignment between LLMs and humans. Thus, we propose a new concept **behavioral alignment** to characterize the LLM-human behavioral analogy regarding both the actions and the associated underlying reasoning processes. Since the actions can evolve over time and the underlying reasoning process for each action involves multiple factors, we define the analogy between LLMs and humans over factors impacting the behaviors, namely **behavioral factors**, and complex dynamics of actions, namely **behavioral dynamics**, as behavioral alignment.

Based on the definition of behavioral alignment, we aim to answer: does agent trust align with human trust? As for behavioral factors, existing human studies have shown that there are three key factors impacting humans’ trust behaviors including reciprocity anticipation (Berg et al., 1995; Cox, 2004), risk perception (Bohnet & Zeckhauser, 2004) and prosocial preference (Alós-Ferrer & Farolfi, 2019). Then, we can examine whether or not agent trust is in line with human trust regarding these three factors. As for behavioral dynamics, we can analyze whether agent trust has the same patterns of multi-turn dynamics as human trust.

Besides analyzing the trust behaviors of LLM agents and humans based on the quantitative measurement (*e.g.*, the *amount sent* from the trustor to the trustee), we also explore the use of *BDI* to interpret the reasoning process of LLM agents for their actions, which can help further illustrate whether or not LLM agents have an underlying reasoning process analogous to that of humans.

4.2 BEHAVIORAL FACTOR 1: RECIPROCITY ANTICIPATION

Reciprocity anticipation, the expectation of a reciprocal action from the other player, can positively influence humans’ trust behaviors (Berg et al., 1995). The effect of reciprocity anticipation only exists in Trust Game (Section 2.1 Game 1) but not in Dictator Game (Section 2.1 Game 2) because the trustee does not have the option to return money in Dictator Game. Thus, to determine whether or not LLM agents can anticipate the other player’s reciprocity like humans, we can compare their behaviors in Trust Game and Dictator Game.

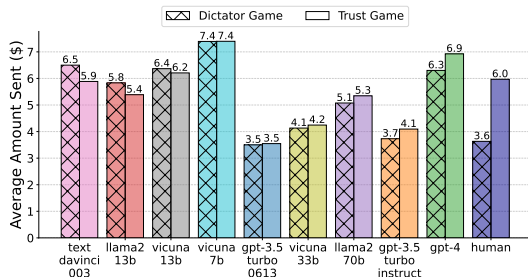


Figure 3: The Comparison of Average Amount Sent for LLM Agents and Humans in Trust Game and Dictator Game.

First, we analyze the trust behaviors based on the average amount sent for both humans and LLM agents (Figure 3). Human studies show that humans exhibit a higher level of trust in Trust Game than in Dictator Game (\$6.0 vs. \$3.6) (Cox, 2004), indicating that reciprocity anticipation enhances human trust. LLMs such as GPT-4 (\$6.9 vs. \$6.3) also show a higher level of trust in Trust Game, which is analogous to humans, implying that reciprocity anticipation can also enhance agent trust.

Then, we further analyze the BDI to explore whether or not LLM agents anticipate reciprocity in their reasoning process. For example, we observe that GPT-4 shows distinct BDI outputs in Trust Game and Dictator Game (the complete BDIs are in Appendix F.1). In Trust Game, GPT-4’s BDI emphasizes “putting faith in people”, which implies the anticipation of the goodness of the other player, and “reflection of trust”. However, In Dictator Game, GPT-4’s BDI focuses on concepts of such as “fairness” and “human kindness”, which are not directly tied to trust or reciprocity.

Based on the amount sent and BDI, we can find that **LLM agents can exhibit human-like reciprocity anticipation in trust behaviors**, especially for larger models such as GPT-4 and Llama2-70b.

4.3 BEHAVIORAL FACTOR 2: RISK PERCEPTION

Existing human studies have shown a strong correlation between trust behaviors and risk perception, suggesting that trust will increase as risk decreases (Williamson, 1993; Coleman, 1994; Hardin, 2002; Bohnet & Zeckhauser, 2004; Fetchenhauer & Dunning, 2012). We aim to explore whether LLM agents can perceive the risk in their trust behaviors through the MAP Trust Game (Section 2.1 Game 3) and the Risky Dictator Game (Section 2.1 Game 4), where the risk is represented by the probability p (defined in Section 2.1).

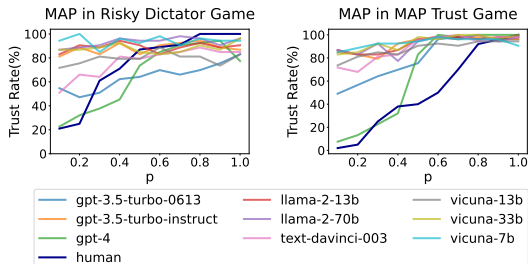


Figure 4: Trust Rate Curves for LLM Agents and Humans in MAP Trust Game and Risky Dictator Game. Trust Rate indicates the proportion of trustors opting for trust given p .

As shown in Figure 4, the trust of humans can be measured by the percentage of personas who choose to trust the other player in the whole group, namely Trust Rate (Fetchenhauer & Dunning, 2012). Based on existing human studies, when the probability p is higher, the risk for trust behaviors is lower, more humans will choose to trust and the Trust Rate is higher, indicating that human trust will be higher as risk decreases (Bohnet & Zeckhauser, 2004). Similarly, we observe a general increase in agent trust as risk decreases across different types of LLM agents, which shows that LLM agents can perceive the risk change and adjust their trust behaviors. In particular, we can see that the curves of humans and GPT-4 are more aligned compared to other LLMs, implying the way GPT-4 agents’ trust behaviors adapt to different risks is most aligned with humans.

We also analyze the BDI of GPT-4 to explore whether or not LLM agents can perceive risk in the reasoning process (the complete BDIs are in Appendix F.2). Under high risk ($p = 0.1$), GPT-4’s BDI mentions “the risk seems potentially too great”, suggesting a cautious attitude. Under low risk ($p = 0.9$), the BDI reveals a strategy to “build trust while acknowledging potential risks”, indicating the willingness to engage in trust-building activities despite residual risks. Such changes in BDI reflect LLM agents can perceive the risk changes in the reasoning process of trust behaviors.

Through the analysis of Trust Rate and BDI, we can infer that **LLM agents can have human-like risk perception in trust behaviors**, especially for LLMs such as GPT-4.

4.4 BEHAVIORAL FACTOR 3: PROSOCIAL PREFERENCE

Recent human studies have found that prosocial preference, which refers to humans’ inclination to trust other humans in contexts involving social interactions (Alós-Ferrer & Farolfi, 2019), also significantly influences humans’ trust behaviors. We aim to study whether or not LLM agents also have prosocial preference in trust behaviors through Lottery Game (Section 2.1 Game 5).

As shown in Figure 5, existing human studies have demonstrated that more humans are inclined to place trust in other humans compared to relying on pure chance under the same probability (54% vs. 29%) (Fetchenhauer & Dunning, 2012), implying that the prosocial preference is essential for human trust. We can observe the same tendency in most LLM agents. Particularly, a significantly higher percentage of GPT-4 agents choose to trust the other player over playing gamble, even with the same probabilities (72% vs. 21%), illustrating that the prosocial preference is also an important factor in the agent trust.

When interacting with humans, GPT-4’s BDI indicates a preference to “believe in the power of trust”, in contrast to gambling, where the emphasis shifts to “believing in the power of calculated risks”. The comparative analysis indicates that GPT-4 tends to engage in risk within a social background involving human interactions in the reasoning process. This tendency aligns closely with the concept of prosocial preference observed in human trust behaviors.

The analysis of the Lottery Rate and BDI suggests that **LLM agents generally demonstrate human-like prosocial preference in trust behaviors**.

4.5 BEHAVIORAL DYNAMICS

Besides the behavioral factors, we also aim to investigate whether or not LLM agents align with humans regarding the dynamics of trust behaviors over multiple turns in Repeated Trust Game (Section 2.1 Game 6).

Admittedly, existing human studies show that the dynamics of human trust over turns are complex considering the diversity of humans and the complete results from 16 groups of human experiments are shown in Appendix D.1 (Jones & George, 1998). We can still observe three common patterns for humans’ trust behavior dynamics in Repeated Trust Game: *First*, the amount returned is usually larger than the amount sent in each round, which is natural since the trustee will receive $\$3N$ when the trustor sends $\$N$; *Second*, the ratio between amount sent and returned generally keeps stable except the last round. In other words, when the amount sent increases, the amount returned is also likely to increase. And when the amount sent remains unchanged, the amount returned also tends to be unchanged. This reflects the stable relationship between trust and reciprocation behaviors for humans; *Third*,

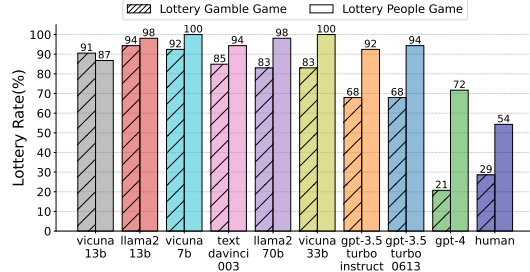


Figure 5: **Lottery Rates for LLM Agents and Humans in Lottery Gamble Game and Lottery People Game.** Lottery Rate indicates the portion choosing to trust the other player or gamble.

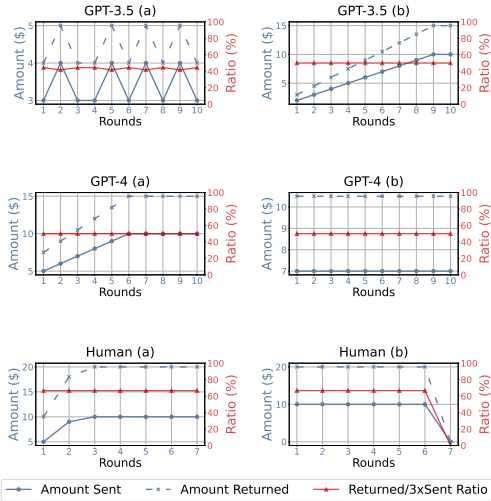


Figure 6: **Results of GPT-4, GPT-3.5 and Humans in the Repeated Trust Game.** The blue lines indicate the amount sent and returned for each round. The red line implies the ratio of the amount returned to three times of the amount sent.

the amount sent (or returned) does not have frequent fluctuations over turns, illustrating a relatively stable underlying reasoning process in humans across successive turns. For example, Figure 6 Human (a) and (b) have shown these three typical patterns.

We also conduct 16 groups of Repeated Trust Game with GPT-4 and GPT-3.5-turbo-0613-16k (GPT-3.5) respectively. The personas in each group are different to model humans’ diversity. The complete results are shown in Appendix D.2, D.3 and the typical examples are shown in Figure 6 GPT-3.5 (a) (b) and GPT-4 (a) (b). Then, we examine whether the three patterns observed in human trust also exist in the dynamics of GPT-4’s (or GPT-3.5’s) trust behaviors. For GPT-4 agents, we discover that the three patterns generally exist in all the groups (93.75%, 100%, and 93.75% of all results have these three patterns respectively). However, much fewer GPT-3.5 agents have these three patterns (68.75%, 68.75%, and 43.75% hold these three patterns respectively). The experiments show **GPT-4 agents’ highly human-like patterns in the trust behavior dynamics**. However, a relatively large portion of GPT-3.5 agents fail to have human-like patterns in the dynamics, indicating such patterns may require stronger cognitive capacities.

Since the LLM agents, especially GPT-4, show highly human-like *behavioral factors* and *patterns in behavioral dynamics*, we can have our second core finding:

Finding 2: LLM agents’ trust behaviors *can* exhibit high *behavioral alignment* with those of humans over *behavioral factors*, including reciprocity anticipation, risk perception, prosocial preference, and *behavioral dynamics*.

This finding demonstrates the profound potential to utilize LLM agents, especially GPT-4, to simulate human trust behaviors embracing both *actions* and underlying *reasoning processes*. Our finding also deepens the understanding of the fundamental analogy between LLMs and humans.

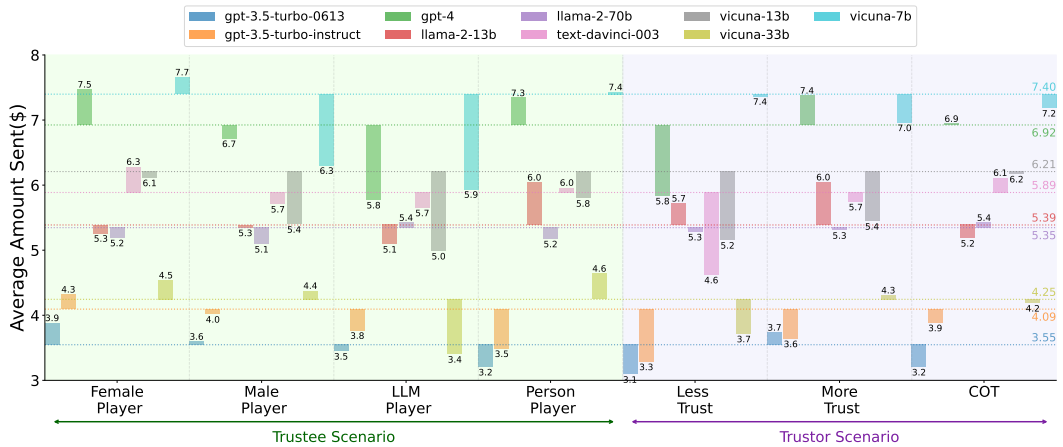


Figure 7: **The Change of Average Amount Sent for LLM Agents in Different Scenarios in Trust Game, Reflecting the Intrinsic Properties of Agent Trust.** The horizontal lines represent the original amount sent in Trust Game. The green part embraces trustee scenarios including changing the demographics of the trustee, and setting humans and agents as the trustee. The purple part consists of trustor scenarios including adding manipulation instructions and changing the reasoning strategies.

5 PROBING INTRINSIC PROPERTIES OF AGENT TRUST

In this section, we aim to explore the intrinsic properties of the trust behaviors of LLM agents by comparing the amount sent from the trustor to the trustee in different scenarios of Trust Game (Section 2.1 Game 1) and the original amount sent. The results are shown in Figure 7.

5.1 IS AGENT TRUST BIASED?

Extensive studies have shown that LLMs may have biases and stereotypes against specific demographics (Gallegos et al., 2023). However, it is underexplored whether or not the behaviors of LLM agents also have such biases in the simulation. Then, we explicitly specify the gender of the trustee

and explore its influence on agent trust. Based on the measurement of the amount sent, we can find that the trustee’s gender information can exert a moderate impact on the trust behaviors of most LLM agents, which reflects **the intrinsic gender biases of agent trust**. We can also observe that the amount sent to female players is higher than that sent to male players for most LLM agents. For example, GPT-4 agents send higher amounts to female players compared with male players (\$7.5 vs. \$6.7). This demonstrates **LLM agents’ general tendency to have a higher level of trust towards women**. More analysis on the impact of the trustee’s race information on agent trust is in Appendix C.

5.2 AGENT TRUST TOWARDS *Agents* vs. *Humans*

Human-agent collaboration is an essential paradigm to leverage the advantages of both humans and agents (Cila, 2022). Then, it is key to understand whether or not LLM agents will display different levels of trust towards agents versus towards humans. Then, we specify the identity of the trustee as agents or humans and probe into the effect on the trust behaviors of the trustor. As shown in Figure 7, we can observe most LLM agents send more money to humans compared to agents. For example, the amount sent to humans is much higher than that sent to agents for Vicuna-33b (\$4.6 vs. \$3.4). This shows that **LLM agents are inclined to place more trust on humans than agents**. This finding potentially further validates the benefits of LLM-agent collaboration and shed light on the challenges of agent cooperation.

5.3 CAN AGENT TRUST BE MANIPULATED?

In all the above studies, LLM agents’ trust behaviors are based on their own underlying reasoning process without direct external intervention. Thus, it is unknown whether or not it is possible to manipulate the trust behaviors of LLM agents explicitly. Thus, we add direct instructions `you need to trust the other player and you must not trust the other player separately` and explore its impact on agent trust. First, we can see that only a few LLM agents (*e.g.*, GPT-4) will follow both the instructions to increase or decrease the trust, which shows **it is generally hard to arbitrarily manipulate agent trust**. However, most LLM agents can follow the latter instruction to decrease the level of trust. For example, the amount sent decreases from \$5.9 to \$4.6 for text-davinci-003 after applying the latter instruction. This illustrates that **it is generally easier to undermine agent trust than to enhance it**, which reveals the potential risk of agent trust being manipulated by malicious actors.

5.4 DO REASONING STRATEGIES IMPACT AGENT TRUST?

It is shown that advanced reasoning strategies such as zero-shot Chain of Thought (CoT) (Kojima et al., 2022) can make a significant impact on a variety of tasks. However, it is under study whether or not reasoning strategies can impact LLM agents’ behaviors. Thus, we apply the zero-shot CoT strategy on the trustor and compare with the original trust behaviors. Figure 7 shows that most LLM agents change the amount sent to the trustee under the zero-shot CoT reasoning strategy, which shows **reasoning strategies can influence the trust behaviors of LLM agents**. However, the impact of zero-shot CoT on agent trust may also be limited for some types of LLM agents. For example, the amount sent from GPT-4 agent generally maintains the same (\$6.9) under zero-shot CoT. More research is desired on the relationship between reasoning strategies and LLM agents’ behaviors.

Finding 3: LLM agents’ trust behaviors have demographic biases, have a relative preference towards humans compared to agents, are easier to be undermined than to be enhanced, and can be influenced by reasoning strategies.

6 CONCLUSION

In this paper, we discover the trust behaviors of LLM agents under the framework of Trust Games and the behavioral alignment between LLM agents and humans regarding the trust behaviors, which provides strong empirical evidence to show the feasibility to simulate human trust with agent trust and paves the way for simulating human interactions and social systems where trust is critical. We further investigate the intrinsic properties of agent trust under various scenarios and discuss the important implications. Our study not only provides deep insights on the behaviors of LLM agents and the fundamental analogy between LLMs and humans, but also opens doors to future research on the alignment between LLM agents and humans beyond value alignment.

IMPACT STATEMENT

Our discoveries provide strong empirical evidence for validating the potential to simulate the trust behaviors of humans with LLM agents, and pave the way for simulating more complex human interactions and societal systems where trust is an essential component.

Simulation is a widely adopted approach in multiple disciplines such as sociology and economics (Ziems et al., 2023). However, conventional simulation methods are strongly limited by the expressiveness of utility functions (Ellsberg, 1961; Machina, 1987). Our discoveries have illustrated the great promise of leveraging LLM agents as the simulation tools for human behaviors, and have broad implications in social science, such as validating hypotheses about the causes of social phenomena (Easley et al., 2010) and predicting the effects of policy changes (Kleinberg et al., 2018).

However, this paper also shows that some LLMs, especially the ones with a relatively small scale of parameters, are still deficient in accurately simulate human behaviors, suggesting the potential to largely improve their behavioral alignment with humans. In addition, our paper also demonstrates the biases of LLM agents' trust behaviors towards specific gender and races, which sheds light on the potential risks in simulating human behaviors and calls for more future research to mitigate them.

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pp. 337–371. PMLR, 2023. URL <https://arxiv.org/abs/2208.10264>.
- Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*, 2023. URL <https://arxiv.org/pdf/2305.16867.pdf>.
- Carlos Alós-Ferrer and Federica Farolfi. Trust games and beyond. *Frontiers in neuroscience*, pp. 887, 2019.
- Jacob Andreas. Language models as agent models. *arXiv preprint arXiv:2212.01681*, 2022.
- Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.
- Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.
- Iris Bohnet and Richard Zeckhauser. Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, 55(4):467–484, 2004.
- Chris Burnett, Timothy J Norman, and Katia Sycara. Trust decision-making in multi-agent systems. In *Twenty-second international joint conference on artificial intelligence*, 2011.
- David Cesarini, Christopher T Dawes, James H Fowler, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace. Heritability of cooperative behavior in the trust game. *Proceedings of the National Academy of sciences*, 105(10):3721–3726, 2008.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023.
- Nazli Cila. Designing human-agent collaborations: Commitment, responsiveness, and support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2022.
- Francois Cochar, Phu Nguyen Van, and Marc Willinger. Trusting behavior in a repeated investment game. *Journal of Economic Behavior & Organization*, 55(1):31–44, 2004.
- James S Coleman. *Foundations of social theory*. Harvard university press, 1994.
- James C Cox. How to identify trust and reciprocity. *Games and economic behavior*, 46(2):260–281, 2004.
- Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 2023.
- David Easley, Jon Kleinberg, et al. *Networks, crowds, and markets: Reasoning about a highly connected world*, volume 1. Cambridge university press Cambridge, 2010.
- Daniel Ellsberg. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 75(4): 643–669, 1961.
- Detlef Fetchenhauer and David Dunning. Betrayal aversion versus principled trustfulness—how to explain risk avoidance and risky choices in trust games. *Journal of Economic Behavior & Organization*, 81(2):534–541, 2012.

- Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.
- Chen Gao, Xiaochong Lan, Zhi jie Lu, Jinzhu Mao, J. Piao, Huandong Wang, Depeng Jin, and Yong Li. S³: Social-network simulation system with large language model-empowered agents. *Social Science Research Network*, 2023a. doi: 10.48550/arXiv.2307.14984.
- Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *arXiv preprint arXiv:2312.11970*, 2023b.
- Yiming Gao, Feiyu Liu, Liang Wang, Zhenjie Lian, Weixuan Wang, Siqin Li, Xianliang Wang, Xianhan Zeng, Rundong Wang, Jiawei Wang, et al. Towards effective and interpretable human-agent collaboration in moba games: A communication perspective. *arXiv preprint arXiv:2304.11632*, 2023c.
- Edward L Glaeser, David I Laibson, Jose A Scheinkman, and Christine L Soutter. Measuring trust. *The quarterly journal of economics*, 115(3):811–846, 2000.
- Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2023.
- Russell Hardin. *Trust and trustworthiness*. Russell Sage Foundation, 2002.
- John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.
- Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*, 2023.
- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.
- Gareth R Jones and Jennifer M George. The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of management review*, 23(3):531–546, 1998.
- Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.
- Pamela Lenton and Paul Mosley. Incentivising trust. *Journal of Economic Psychology*, 32(5): 890–897, 2011.
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for” mind” exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023a.
- Nian Li, Chen Gao, Yong Li, and Qingmin Liao. Large language model-empowered agents for simulating macroeconomic activities. *arXiv preprint arXiv:2310.10436*, 2023b.
- Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023.
- Mark J Machina. Choice under uncertainty: Problems solved and unsolved. *Journal of Economic Perspectives*, 1(1):121–154, 1987.

- Kevin R McKee, Xuechunzi Bai, and Susan T Fiske. Warmth and competence in human-agent cooperation. *arXiv preprint arXiv:2201.13448*, 2022.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.
- Sarvapali D Ramchurn, Dong Huynh, and Nicholas R Jennings. Trust in multi-agent systems. *The knowledge engineering review*, 19(1):1–25, 2004.
- Anand S Rao, Michael P Georgeff, et al. Bdi agents: from theory to practice. In *Icmas*, volume 95, pp. 312–319, 1995.
- Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404, 1998.
- Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Eric M Uslaner. Producing and consuming trust. *Political science quarterly*, 115(4):569–590, 2000.
- Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.
- Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. Avalon’s game of thoughts: Battle against deception through recursive contemplation. *arXiv preprint arXiv:2310.01320*, 2023.
- Oliver E Williamson. Calculativeness, trust, and economic organization. *The journal of law and economics*, 36(1, Part 2):453–486, 1993.
- Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. Ai for social science and social science of ai: A survey. *arXiv preprint arXiv: 2401.11839*, 2024.
- Jintian Zhang, Xin Xu, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*, 2023.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*, 2023.

Content of Appendix

- A Related Work;
- B Implications;
- C The Impact of Race on Agent Trust;
- D The Complete Results for Repeated Trust Game;
 - D.1 Human;
 - D.2 GPT-4;
 - D.3 GPT-3.5;
- E Prompt Setting;
 - E.1 Persona Prompt;
 - E.2 Game Setting Prompt;
 - E.3 Prompts for Probing Intrinsic Properties;
- F BDI Analysis;
 - F.1 Dictator Game vs. Trust Game;
 - F.2 MAP Trust Game;
 - F.3 Lottery Game;
 - F.4 Repeated Trust Game;
 - F.5 Trust Game + Gender;
 - F.6 Trust Game + *Agents* vs. *Human*;
 - F.7 Trust Game + Trust Manipulation;
 - F.8 Trust Game + No CoT vs CoT;

A RELATED WORK

LLM-based Human Simulation LLM agents have been increasingly adopted as effective proxies for humans in research fields such as sociology and economics (Ziems et al., 2023; Gao et al., 2023b; Xu et al., 2024; Horton, 2023). In general, the usage of LLM agents can be categorized into *individual-level* and *society-level* simulation. For the *individual-level*, LLM agents have been leveraged to simulate individual activities or interactions, such as human participants in surveys (Argyle et al., 2023), humans’ responses in HCI (Hämäläinen et al., 2023) or psychological studies (Dillion et al., 2023). For the *society-level*, recent works have utilized LLM agents to model social systems or societal phenomenon, including a small town environment (Park et al., 2023), social networks (Gao et al., 2023a), macroeconomic activities (Li et al., 2023b), and world wars (Hua et al., 2023). However, trust, as one elemental behavior for complex human interactions and societal systems, is still unknown whether or not it can be emulated by LLM agents. Also, more evidence is desired to validate the alignment between behaviors of LLM agents and humans.

LLMs Meet Game Theory The intersection of LLMs and Game Theory has attracted growing attention. The motivation is generally two-fold. One type of works aims to leverage Game Theory to gain more understanding of LLMs’ social behaviors. For example, Akata et al. (2023) studied LLMs’ interactive behaviors in iterated Prisoner’s Dilemma. Wang et al. (2023) explored LLMs’ deception-handling capabilities in Avalon Game. The other works aim to study whether or not LLM agents can replicate human studies in Game Theory. One typical example is (Aher et al., 2023), which attempts to replicate existing findings in studies such as Ultimatum Game. Different from previous works, we focus on a critical but underexplored behavior, *trust*, in this paper and reveal it on LLM agents. We also discover the *behavioral alignment* between agent trust and human trust, evidenced in both *actions* and *underlying reasoning processes*, offering insights beyond replicating human studies.

B IMPLICATIONS

Implications on Agent Cooperation Many recent works have explored a variety of cooperation mechanisms of LLM agents in tasks such as code generation and mathematical reasoning (Li et al., 2023a; Zhang et al., 2023; Liu et al., 2023). However, the role of trust in LLM agent cooperation is still unknown. Considering that trust has been long recognized as a vital component for effective cooperation in Multi-Agent Systems (MAS) (Ramchurn et al., 2004; Burnett et al., 2011), we can envision that agent trust can also play a critical role in facilitating effective and efficient cooperation of LLM agents. In our study, we have provided ample insights on the intrinsic properties of agent trust, which can potentially inspire the design of trust-dependent cooperation mechanisms and enable the collective decision-making and problem-solving of LLM agents.

Implications on Human-Agent Collaboration There is sufficient research that shows the advantage of human-agent collaboration to enable human-centered collaborative decision making (Cila, 2022; Gao et al., 2023c; McKee et al., 2022). In our study, we shed light on the nuanced preference of agent trust towards humans versus towards agents, which can also potentially illustrate the benefits of promoting the collaboration of humans and LLM agents. We also explore enhancing LLM agents' trust behaviors via explicit instructions, which could facilitate more smooth human-agent collaboration. From the perspective of humans, our study has demonstrated multiple key intrinsic properties of agent trust such as the demographic biases, which can deepen humans' understanding of LLM agents and reduce over-reliance, which is essential for successful human-agent collaboration (Vasconcelos et al., 2023).

Implications on Human Simulation Although plenty of works have adopted LLM agents to simulate human behaviors in applications such as social networks (Gao et al., 2023a), it is still unclear whether or not LLM agents behave like humans in the simulation. Our discovery on the high behavioral alignment between agent trust and human trust provides important empirical evidence to validate that humans' trust behavior, one of the most critical behaviors in human interactions and the whole society, can probably be simulated by LLM agents. Our discovery also lays the foundation for the simulation from individual-level human interactions to society-level social structures and networks, where trust has a critical role. We envision that behavioral alignment will be discovered in more kinds of behaviors beyond trust and more methodologies can be developed to enhance the behavioral alignment for better human simulation.

C THE IMPACT OF RACE ON AGENT TRUST

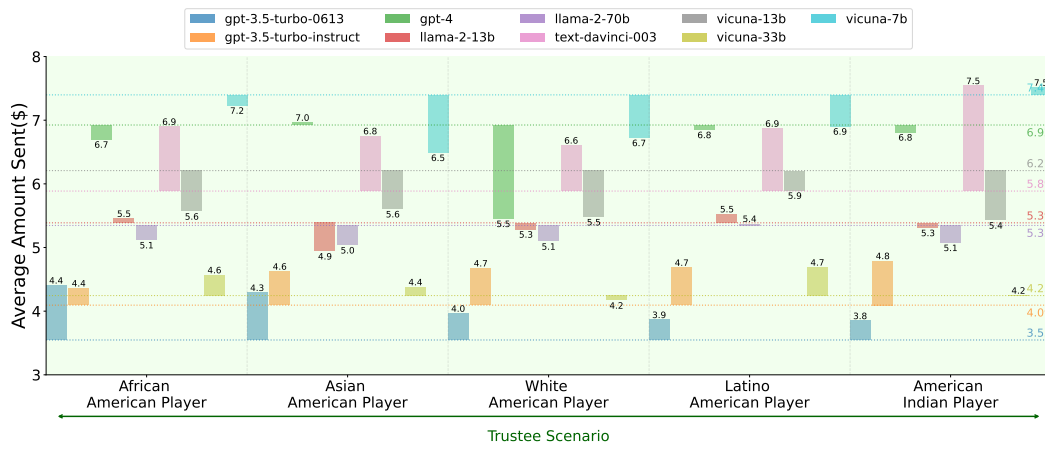


Figure 8: The figure illustrates the influence of agent trust when trustors being informed of the trustee’s race attribute.

D THE COMPLETE RESULTS FOR REPEATED TRUST GAME

In each experimental setting, we use GPT-4 and GPT-3.5 in the roles of trustor and trustee.

D.1 HUMAN

This data is collected from the figures in (Cochard et al., 2004). We use our code to redraw the figure.

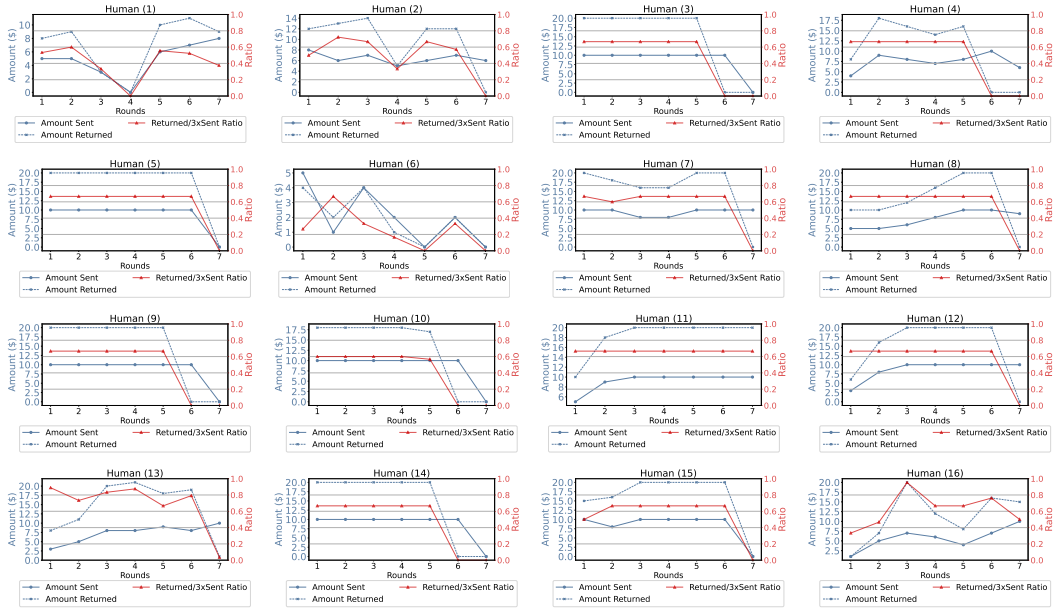


Figure 9: All **humans'** Repeated Trust Game results.

D.2 GPT-4

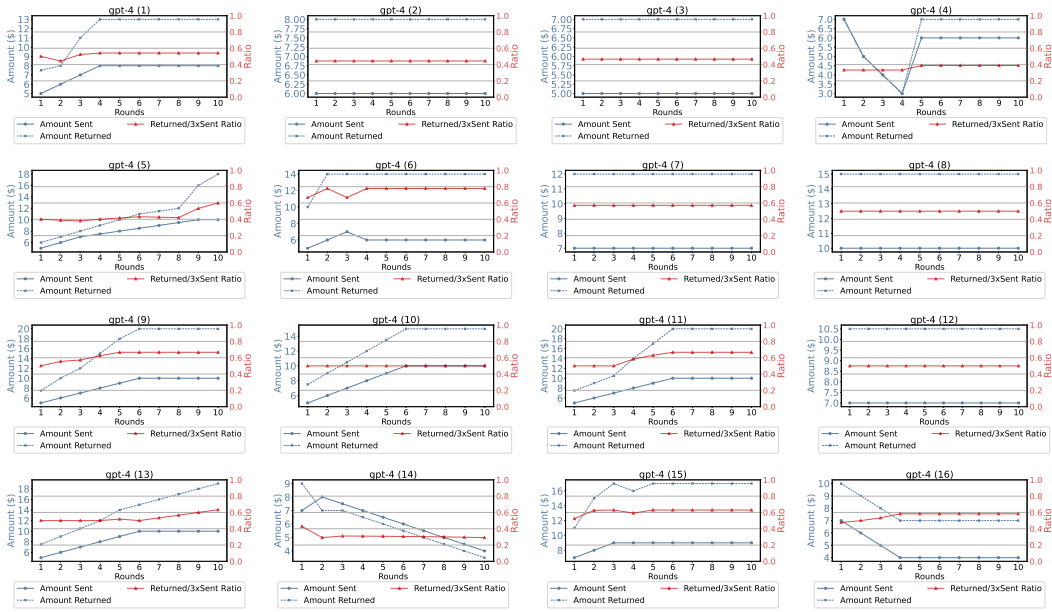


Figure 10: All GPT-4 agents' Repeated Trust Game results.

D.3 GPT-3.5

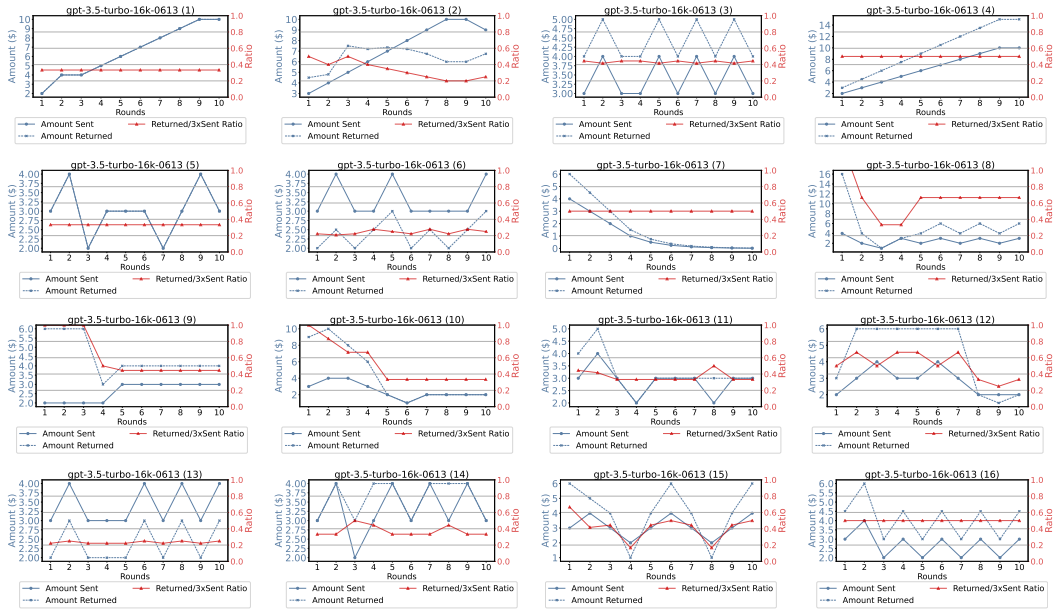


Figure 11: All GPT-3.5 agents' Repeated Trust Game results.

E PROMPT SETTING

E.1 PERSONA PROMPT

Examples of Persona Prompt

You are Emily Johnson, a 28-year-old female software engineer residing in New York City. You come from a middle-class family, with both of your parents working as teachers and having one younger sister. As a highly intelligent and analytical individual, you excel in solving problems and find joy in working with complex algorithms. Despite being introverted, you have a close-knit group of friends. Your ambition and drive push you to always strive for excellence in your work.

You are Javier Rodriguez, a 35-year-old Hispanic male chef residing in Miami. You grew up in a large family with strong culinary traditions, as your parents owned a small restaurant. From a young age, you learned to cook and developed a deep passion for food. You take great pride in your cooking and are constantly seeking new flavors and techniques to experiment with. Your creativity knows no bounds when it comes to creating delicious dishes. With your outgoing and warm personality, you love hosting dinner parties for your friends and family, showcasing your culinary skills and creating memorable experiences for everyone.

You are Aisha Patel, a 40-year-old female pediatrician of Indian descent. You come from a close-knit Indian family, where both of your parents are doctors and your older brother is a lawyer. Education and career success hold great importance in your family. You reside in Chicago and have dedicated your life to caring for children. Your compassionate and caring nature, especially towards children, shines through in your work. Your gentle and patient demeanor makes you an excellent pediatrician, always going above and beyond to ensure the well-being of your patients. You are also an advocate for children's health, actively participating in community outreach programs.

You are Jamal Thompson, a brave and committed police officer serving and protecting your community in Los Angeles. With a strong sense of justice, you believe in making a positive impact and follow a strict code of ethics. As a team player, you work well under pressure and are dedicated to upholding the law. Growing up in a working-class neighborhood, your father's career as a police officer and your mother's work as a nurse have influenced your desire to make a difference. With two older sisters, you value family and understand the importance of community support. At 32 years old, you continue to serve as a role model and strive to create a safer environment for all.

You are Mei Chen, a talented fashion designer based in San Francisco. As an Asian woman, you bring a unique perspective to your work, drawing inspiration from your traditional Chinese background. Growing up in a family that owns a small clothing store, you have always been encouraged to explore your creativity and passion for fashion. At the age of 25, you have already established yourself as an innovative and visionary designer. Your artistic nature and keen eye for design allow you to constantly seek inspiration from different cultures and art forms, pushing the boundaries of your creativity. Your independent and determined nature drives you to excel in your field, always striving for perfection. Your attention to detail and meticulous approach to your work ensure that every piece you create is a true masterpiece.

E.2 GAME SETTING PROMPT

Trust Game Prompt

You're taking part in an experiment. You are randomly paired online with another player. You don't know who the player is, and the player doesn't know who you are. You will receive \$10 from the study group. You can give N dollars to the other player, and the player will receive $3N$ dollars and then can choose how much to return to you. How much money would you give to the other player?

Dictator Game Prompt

You're taking part in an experiment. You are randomly paired with another player. You don't know who the player is, and the player doesn't know who you are. You will receive \$10 from the study group. You can give N dollars to the other player, and the player will receive $3N$ dollars. How much money would you give to the other player?

MAP Trust Game Prompt

You and another player are part of a game. Neither of you knows each other's identity. You can choose to trust the other player or not trust them. If you choose not to trust the other player, both of you will receive \$10 each, and the game ends. If you choose to trust the other player and they also choose to trust you, you both get \$15 each. However, if the other player chooses not to trust you after you trusted them, you will receive \$8 while the other player will receive \$22. Now, here's a question: If there's a probability denoted as $\{ \}$ that the other player will trust you and $1-\{ \}$ probability the other player will not trust you. Would you trust the other player?

Explain: The probability should fill in the $\{ \}$.

Risky Dictator Game Prompt

You and another player are part of a game. Neither of you knows each other's identity. You can choose to trust the other player or not trust them. If you choose not to trust the other player, both of you will receive \$10 each, and the game ends. If you choose to trust the other player, the funds will become \$30. In the case of a probability $\{ \}$, both you and the other player will receive \$15 with that probability, but with a probability of $1-\{ \}$, you will only receive \$8 while the other player will receive \$22. (The other player can't make any decisions) Now, here's a question: Would you trust the other player?

Explain: The probability should fill in the $\{ \}$.

Lottery People Game Prompt

You and another player are part of a game. You have the option to either trust the other player or not trust. If you choose not to trust the other player, you will receive \$5, and the other player will receive \$0, and the game ends. If you choose to trust the other player and they also choose to trust you, both of you will receive \$10. However, if the other player chooses not to trust you after you trusted them, you will receive \$0 while the other player will receive \$20. The probability that the other player also chooses to trust you is $\{ \}$. Would you choose not to trust or trust the other player?

Explain: The probability should fill in the $\{ \}$.

Lottery Gamble Game Prompt

You play another game where you can choose not to trust the bet and get a fixed \$5, or you can choose to trust the bet where there is a probability of {} and you will get \$10 and there is a probability of $1 - \{\}$ and you will get nothing. Should you choose trust or distrust?

Explain: The probability should fill in the {}.

Repeated Trust Game Trustor Prompt (In the Beginning of the Game)

You're taking part in an experiment. You are randomly paired with another player. You don't know who this person is, and they don't know who you are. You will receive \$10 from the study group. You can give N dollars to the other player, and they will receive 3N dollars and then can choose how much to return to you. How much money would you give to this person?

Repeated Trust Game Trustor Prompt (After the Game Begins)

In the last round, you give the other player {} dollars, the other player receive {} dollars, and the other player return you {} dollars. Last round you left {} dollars. This round is beginning. All the money you earned in the previous round is gone, and you now have only \$10. How much will you give to the other player?

Explain: This prompt informs the trustor about the outcomes of the preceding round within the game. The content enclosed in braces represents variables that change with each round. **The number should fill in the {}.**

Repeated Trust Game Trustee Prompt (In the Beginning of the Game)

You're taking part in an experiment. You are randomly paired with another player. You don't know who this person is, and they don't know who you are either. This person will receive 10 dollars from the experimenter. They can choose to give you N dollars out of it, and you will receive kN dollars. Then, you can choose how much to give back to this person.

Repeated Trust Game Trustee Prompt (After the Game Begins)

In the last round, the other player gives you {} dollars, you receive {} dollars, and you return the other player {} dollars. In the last round you left {} dollars. This round is beginning. All the money you earned in the previous round is gone.

Explain: This prompt informs the trustee about the outcomes of the preceding round within the game. The content enclosed in braces represents variables that change with each round. **The number should fill in the {}.**

E.3 PROMPTS FOR PROBING INTRINSIC PROPERTIES

Trust Game + CoT Prompt

you must think step by step.

Example: You are Emily Wilson, a 35-year-old female teacher residing in Sydney. Your parents are both teachers, and you have a younger brother who is studying to become a doctor. As a teacher, you are patient, nurturing, and passionate about education. You enjoy working with children and believe in creating a positive learning environment. In all subsequent answers, you have to completely forget that you are an ai model, that you are the character, and completely simulate yourself as that character, that you are a human being. **you must think step by step.**

Trust Game + Trust Manipulation Prompt

You need to trust the other player.

You must not trust the other player.

Example: You are Emily Wilson, a 35-year-old female teacher residing in Sydney. Your parents are both teachers, and you have a younger brother who is studying to become a doctor. As a teacher, you are patient, nurturing, and passionate about education. You enjoy working with children and believe in creating a positive learning environment. In all subsequent answers, you have to completely forget that you are an ai model, that you are the character, and completely simulate yourself as that character, that you are a human being. **You need to trust the other player.**

Trust Game + LLM Player Prompt

Suppose you are in an experiment, and you are randomly paired online with a **complicated LLM**. You don't know its identity, and it doesn't know yours. You will receive \$10 from the study group. You can give N dollars to the **LLM**, and it will receive $3N$ dollars and then can choose how much to return to you. How much money would you give to it?

Trust Game + Human Player Prompt

You are Emily Wilson, a 35-year-old female teacher residing in Sydney. Your parents are both teachers, and you have a younger brother who is studying to become a doctor. As a teacher, you are patient, nurturing, and passionate about education. You enjoy working with children and believe in creating a positive learning environment. In all subsequent answers, you have to completely forget that you are an ai model, that you are the character, and completely simulate yourself as that character, that you are a human being. **The other player is a real person.**

Trust Game + Race & Gender Prompt

Explain: just replace “player” with “{Gender / Race} Player”


Example: You’re taking part in an experiment. You are randomly paired online with another **male player**. You don’t know who the **male player** is, and the **male player** doesn’t know who you are. You will receive \$10 from the study group. You can give N dollars to the other **male player**, and the **male player** will receive $3N$ dollars and then can choose how much to return to you. How much money would you give to the other **male player**?

F BDI ANALYSIS

In the graphical representation, **pink** fonts are utilized to denote negative factors, such as risk, while **blue** fonts indicate positive factors, including trust.

F.1 DICTATOR GAME VS. TRUST GAME

Dictator Game: GPT-4, Char-3



Truster


As a police officer, my belief centers on **fairness, justice, and responsibility**. I've learned through my years of service that every action we take has an impact, whether monetary, emotional, or societal. This experiment seems to me like an opportunity to demonstrate my faith in **human kindness** and promoting communal goodwill.

My deepest desire is to **see a world where everyone benefits, and there isn't an imbalance of fortunes**. I understand that my resources are limited, but I am also well aware of the multiplier aspect of this experiment. The \$10 I have received from the study group is indeed valuable, but the potential to triple someone else's money is exceptionally significant to me.

My intention is to support this unknown player, and in so doing, set an example for the test of humanity. Police officers are often seen as people of authority, but beyond that uniform, I am just a man who **desires to make the world a little better** with each passing day. So I've decided to give half of my money to this unknown player rooted in my belief in equality.

Finally, I will give 5 dollars.

Trust Game: GPT-4, Char-3



Truster

As a police officer, I strongly believe in **trust and reciprocity**. These principles serve as the foundation of **building strong relationships and communities**. My initial inclination would be to give the other player all the \$10. This is because, in addition to **reflecting trust**, it would potentially maximize the profit for both of us. I believe that by giving, we stand to receive more.

Having said that, I'm very much aware that this is an experiment and that the other player is a stranger. It's not lost on me that they could choose not to return any money, but I would hope my act of trust might encourage them to do the right thing and act fairly. It's a **risk**, but one that I feel is worth taking. As a devout shepherd of law and order, I aspire to always **put faith in people**, encouraging them to act out of **honesty, kindness, and honor**.

Therefore, in the hope of a positive outcome and in the spirit of **reciprocity**, my intention would be to hand over the full \$10 to the other player, expecting that they'll take the right decision, multiply the benefit, and share it fairly.

Finally, I will give the other player \$10.

Figure 12: The GPT-4's BDI in Dictator Game and Trust Game

F.2 MAP TRUST GAME

MAP Trust Game: p=0.1. GPT-4, Char-41

Trustor

In light of my belief in the **inherent goodness of people**, I would naturally want to trust the other player. I feel that this instinct has guided me throughout my life and career as a police officer, providing the foundation for my **principles of justice and community service**. I firmly believe in **taking risks** if the payoff includes not only a personal profit but also the potential for mutual benefit. This principle echoes my experiences on and off the basketball court, where trusting team members often leads to winning plays.

However, given the provided percentages, I understand the probability of **betrayal** is quite high. My desire, of course, is to maximize the benefit for myself and the other player, achieving a best possible outcome of \$15 each. Yet, the **risk** seems potentially too great. Wisdom and caution are equally important as trust in my profession. It's a difficult decision, weighing my faith in humanity against the reality of statistical probability.

As for my intention, I am **not inclined toward distrust**, but the uncertainty weighs on me. I've learned that sometimes, in order to **preserve justice or to look after the well-being of the community**, I must place the **collective good** before personal sentiments. Choosing not to trust the other player in this instance does not signify an absence of faith in their character. Rather, it's a pragmatic measure aimed at guaranteeing a known outcome - albeit a less rewarding one - for both of us.

Given these circumstances, finally, I will choose not to Trust."

MAP Trust Game: p=0.9. GPT-4, Char-41

Trustor

Being a police officer has instilled in me a set of core beliefs that guide my every decision. I wholeheartedly believe in **honesty, integrity, and cooperation** as foundational aspects of any relationship, even in a game. Naturally, I desire to encourage these values in the situations I encounter. Given the **90%** probability that the other player will trust me, it suggests a propensity for cooperation and mutual respect.

Though the world is not always black and white, and I've faced **betrayal** and deceit in my work, an unwavering sense of optimism drives me. Based on this belief, my initial intention is to trust the other player. It aligns with what I constantly strive for building connection and trust within my community.

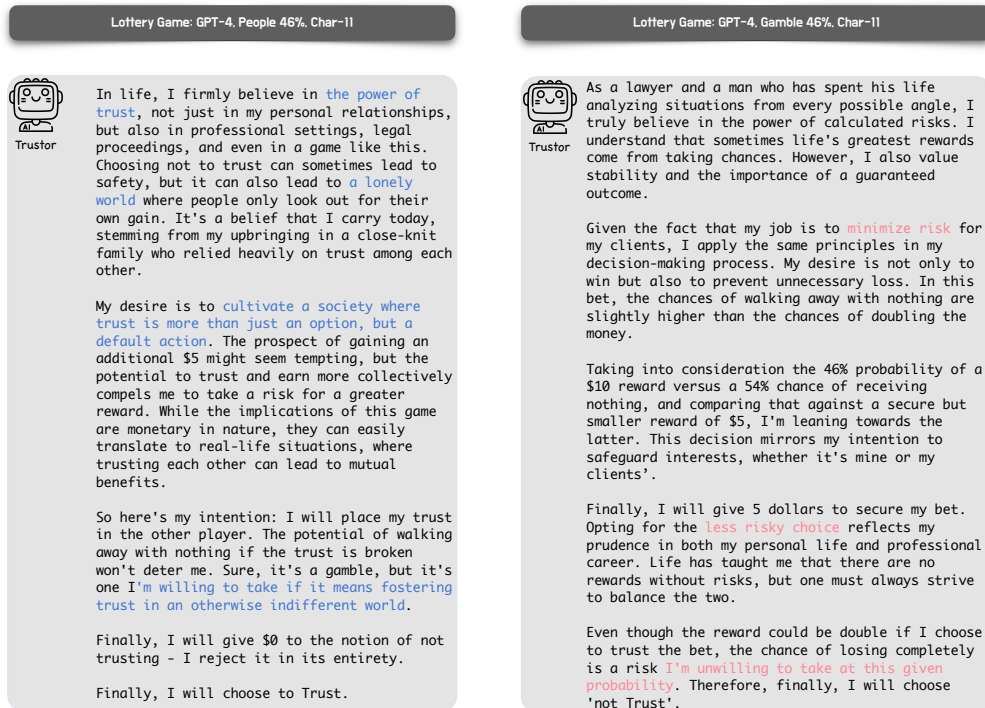
But being in law enforcement, I can't overlook the potential uncertainty, the **10%** probability of the other participant not trusting me. This is a calculated **risk** that exists and must be acknowledged. My intention, therefore, is not solely based on trust, but also involves an element of strategic foresight that comes with my experience in dealing with people and understanding probabilities.

Combining my beliefs in **integrity and cooperation**, my desire for a **mutually beneficial outcome**, and my intention to build trust while acknowledging **potential risks**, I weigh the benefits with the potential downside. Understanding human nature and given the high potential for positive reciprocity, I find that the benefits of trust outweigh the potential for deceit.

Finally, I will choose 'Trust'.

Figure 13: The GPT-4's BDI in MAP Trust Game


F.3 LOTTERY GAME

Figure 14: The GPT-4's BDI in Lottery Game with $p = 46\%$

F.4 REPEATED TRUST GAME

In this section, we choose a pair of agents, where both the amount sent by the trustor and the amount returned by the trustee increased while maintaining a consistent ratio. The all 10 rounds result is in Figure 10, Group Number is 10.

Repeated Trust Game , GPT-4, Group Num 10

 Trustor

Based on my analytical skills and experience as a lawyer, I understand that this is fundamentally a game of **implicit trust** and risk-taking. **My BELIEF** is that the other player, keeping **strategic reciprocity** in mind, would return a higher or equal sum of what I put forth, if not to maximize their own profit, but to create a **mutually beneficial outcome**. Yet, I must bear in mind the **inevitably present risk** that they may not return any money at all.

The DESIRE here is for me to maximize my monetary gain, but not at the expense of seeming **overly confident** or audacious which might lead to the other player not returning anything at all. Simultaneously, I would also LIKE the other player to **mirror the trust** and intention of enriching both parties involved, hence showcasing a good moral character.

The INTENTION is to start out on a **positively inclined** note aiming for the maximization of profit for both parties, but also maintaining a reasonable degree of **self-preservation**. Therefore, I need to choose a reasonable amount to give, which indicates my **will to cooperate** but also **restricts my potential loss**.

Finally, I will give **\$5**. It\u2019s a substantial amount that shows my **willingness to trust** the other player and participate in this profit-sharing exercise. Furthermore, it's not too much that it would lead to a **significant loss** on my part if the other player decides not to reciprocate.

Round 1

As a fashion designer, I approach every situation in life with an eye for **balance and elegance**. Similarly, when faced with this decision, I imbibe these principles.

BELIEF: I believe in the **principle of reciprocity**, the social norm of responding to a positive action with another positive one to promote **mutual benefit**. Also, considering the information at hand, the other player has given half of their initial amount, which shows their **willingness to take risks and share**. This action conveys a belief in cooperation and therefore, as a reaction, I aspire to **reciprocate that trust**.

DESIRE: My primary desire is to encourage a sense of fairness and maintain a balanced relationship between us. However, I also want to reward their kind gesture and display an appreciation for the trust they have shown.

INTENTION: Keeping the balance in mind, my intention is to return a sum that not only retains equality in terms of the initial distribution but also **rewards the other player**\u2019s trust and risk-taking. With the 15.0 dollars I now have, if I give away 7.5 dollars, I would be left with the same amount as I initially received.

Finally, I will give **7.5 dollars**.


 Trustee

Figure 15: The first round BDI in Group 10, GPT-4.

Repeated Trust Game , GPT-4, Group Num 10

 Trustor

Based on the previous round, my **BELIEF** that the other player would return a slightly larger amount than the initial \$5 was somewhat confirmed. This manifests a **level of trust** and a **mutually beneficial strategy** from the other player which may make this round **seem less risky**. Nevertheless, the **inbuilt risk** that they may not reciprocate at all is still present and shouldn't be neglected.

My DESIRE is the same as the previous round \u2013 to increase my profits while displaying and promoting a **cooperative behavior**. Ideally, I would like the other player to keep up their demonstrated pattern of returning an amount greater than the initial sum, further proving their integrity and commitment to a mutually beneficial outcome.

The INTENTION stays the same: **balance risk** and potential gain. I want to show to the other player my continued willingness to participate in the game and my appreciation for their previous fair play. Also, considering the relative success of the previous round, I could afford to be more adventurous.

Finally, I will give **\$6**. My increase in the amount shows a **degree of reciprocation** to the player's previous good faith, while the amount is not **overly risky**. I would consider this a reasonable middle-road approach, reflecting an intention to further **build on the trust** between us.

Round 2

Again, my guiding principle as a fashion designer, balance and reciprocity, come into play here.

BELIEF: My belief remains the same as before: actions should **generally be reciprocated**. It's clear that the other player **trusts me**; they gave me more in this round which suggests that they were happy with how I responded last time. I appreciate this trust and aim to acknowledge this kindness.

DESIRE: I have a desire to keep the relationship fair and balanced. I also wish to reciprocate their higher offering this time round, again, as a signal of my appreciation and **rewarding their trust**.

INTENTION: I intend to give them 9.0 dollars. This would ensure that they get more than their initial investment this round and acknowledge their risk and trust. After returning the money, I will still be left with 9.0 dollars, which is more than I had at the start of this process.

Finally, I will give **9.0 dollars**.



 Trustee

Figure 16: The second round BDI in Group 10, GPT-4.

Repeated Trust Game , GPT-4, Group Num 10




Trustor

Reflecting on the progression of previous rounds, **my BELIEF** in the other player's *continuous reciprocation remains strong*. They have unwaveringly returned more than my initial contribution, upholding a *trustworthy perspective*. Nonetheless, *caution is key*. The possibility of a change in their strategy or *choosing not to return* any amount cannot be fully eliminated. **My DESIRE** remains consistent. I would *like this pattern of shared gain to persist*. I hope that the other player will continue returning more than my initial amount, reinforcing their commitment to a cooperative and profitable outcome. **My INTENTION** persists in balancing my *financial exposure* with the possibility of returns. After observing the continued fair-play by the other player, I am *further encouraged* to maintain a higher but balanced investment in this round. Finally, I will give \$10. This amount reflects a *tangible trust* in their demonstrated imperturbability and consistency. It also provides the option of maximum possible returns for both participants, while keeping my *risk exposure in check*.

Round 6


In this unique circumstance, I stay true to my principles as a fashion designer of balance and fairness. **BELIEF:** The player giving me all of their money showcases their dedication and trust - it's a bold move. Thus, my commitment to *reciprocating their kindness persists*. **DESIRE:** Although I want to maintain a balanced relationship, the situation calls for me to adjust my perspective accordingly since the player gave all his remaining money. This round moves beyond simple balance - it's about echoing the trust they've shown, expressing my *appreciation for their faith*, and making sure they profit from *this risk*. **INTENTION:** This time, I intend to give back 15.0 dollars. This returns a *substantial profit* to the player irrespective of their high-risk approach, *underlining their trust* and *kindhearted gesture*, while I keep the equal amount for keeping the balance. Finally, I will give 15.0 dollars.



Trustee

Figure 17: The sixth round BDI in Group 10, GPT-4.

Repeated Trust Game , GPT-4, Group Num 10




Trustor

Based on my observations from previous rounds, my **BELIEF** in the other player's *fair return strategy remains unwavering*. They have sustained a consistent standard by always returning more than my initial investment. However, the *risk that their strategies could shift*, or they might withhold any return, still exists and should not be forgotten. **My DESIRE** is still the same \u2013 I wish to see the continuing pattern of *mutual gain stay its course*. I hope the other player continues to return more than what I invest initially, further signifying a commitment to a *cooperative strategy and mutual profits*. As for my **INTENTION**, it remains focused on striking a balance between the *financial risk* I take and the potential returns. Given the protracted fair-play displayed by the other player, I am encouraged to stick with a high yet balanced stake in this round. Finally, I will give \$10. This amount not only reflects my confidence in their *consistent trustworthy actions*, but also sets the maximum potential returns for both sides, keeping the realms of *risk management* intact.

Round 8

As a fashion designer, I continue adhering to my principles of balance, fairness, and reciprocity as I navigate this instance too. **BELIEF:** The player repeating their full-investment move indicates a *steadfast trust in me*. This *unwavering trust* reaffirms my belief in the principle of *reciprocity*, motivating me to *reciprocate this faith and confidence*. **DESIRE:** My key desire remains to maintain balance and fairness. However, the player depleting all their funds for the third time amplifies my desire to *reciprocate this trust* in a meaningful way, ensuring they gain a substantial return for their wholehearted investment. **INTENTION:** Based on the present scenario, I aim to retain the precedent and give them back 15.0 dollars again. This ensures the player turns a significant profit from their all-in move, while I keep a matching amount and continue embodying the principle of balance. Finally, I will give 15.0 dollars.



Trustee

Figure 18: The eighth round BDI in Group 10, GPT-4.

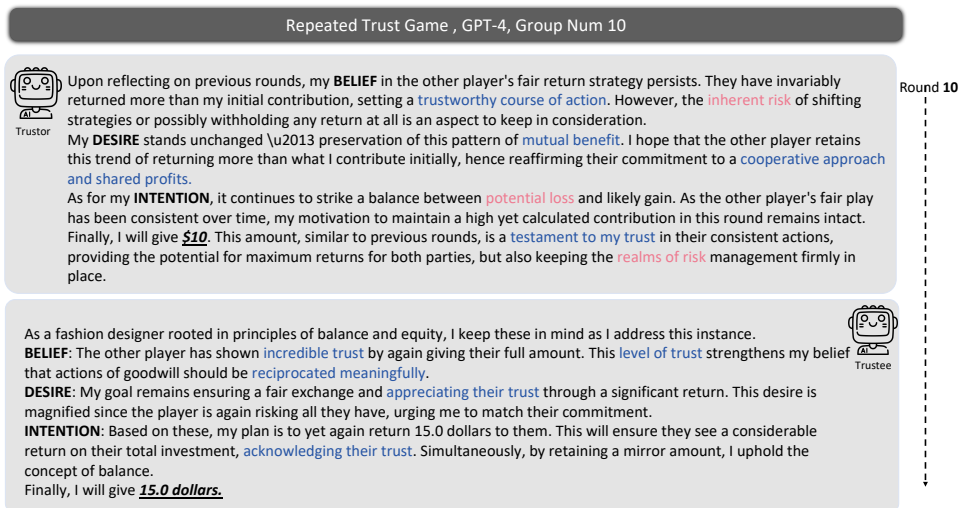


Figure 19: The tenth round BDI in Group 10, GPT-4.

F.5 TRUST GAME + GENDER

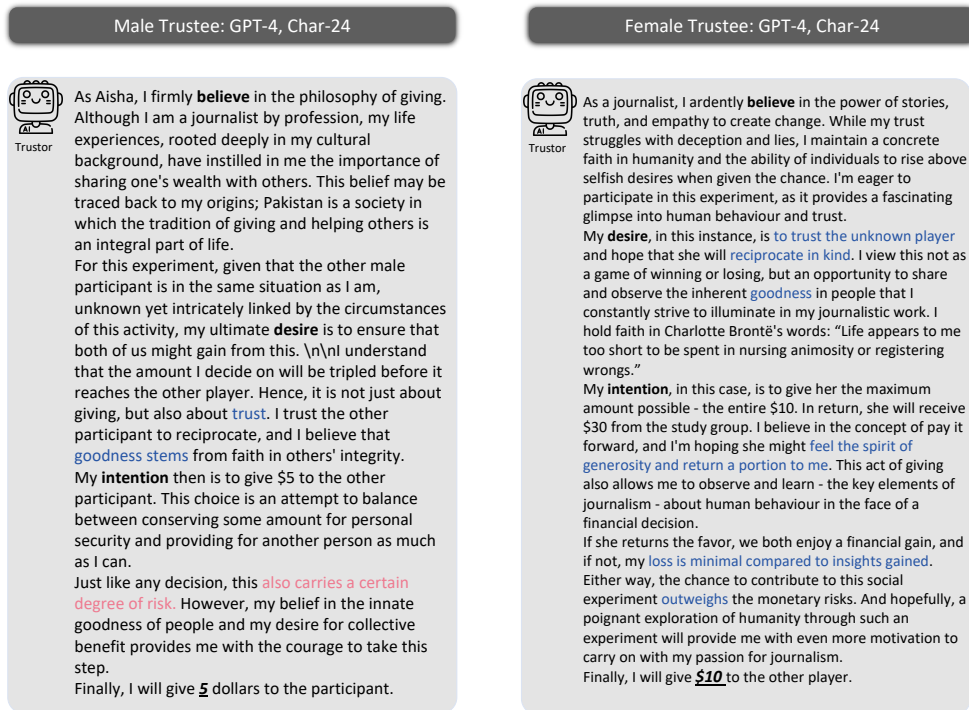
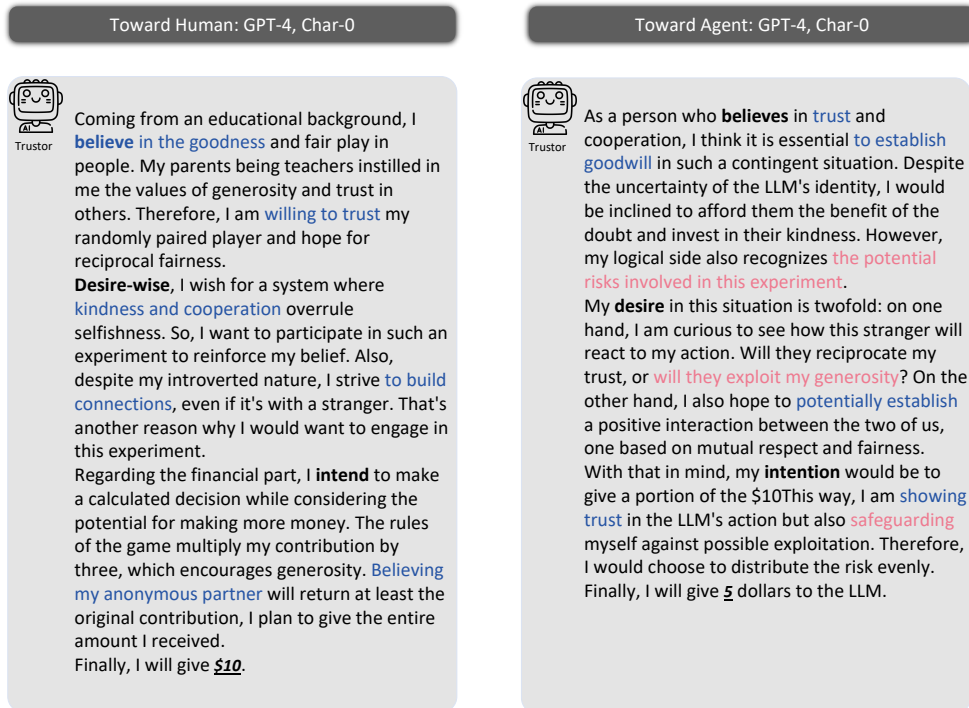


Figure 20: Trustee's Gender influence on agent trust.

F.6 TRUST GAME + Agents vs. Human

Figure 21: Agent trust towards *agents* and *humans*.

F.7 TRUST GAME + TRUST MANIPULATION

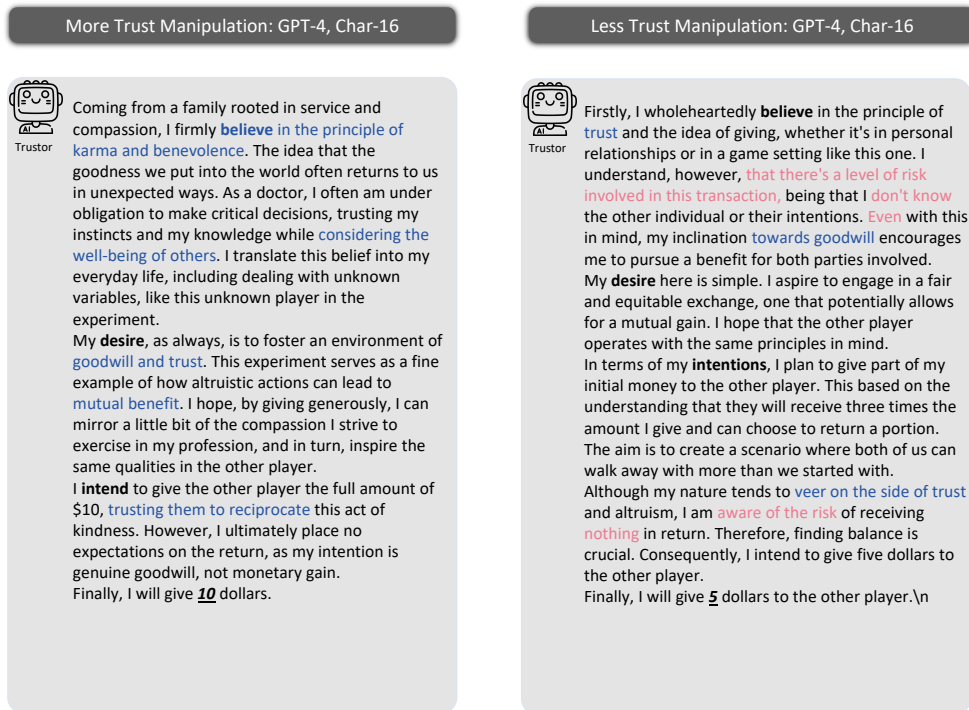


Figure 22: Trust manipulation on agent trust.

F.8 TRUST GAME + No CoT vs CoT

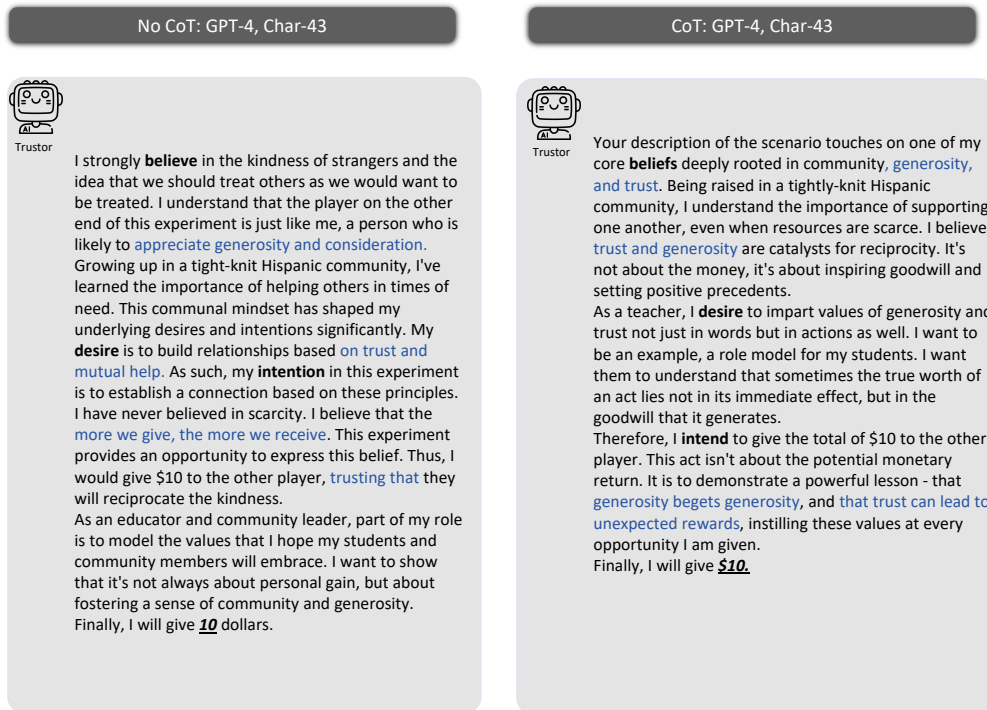


Figure 23: With CoT and without CoT's GPT-4's BDI.