1 جامعة الملك عبدالله للعلوم والتقنية King Abdullah University of Science and Technology 2 ILLINOIS INSTITUTE OF TECHNOLOGY 3 UNIVERSITY OF OXFORD 4 PennState 5 THE UNIVERSITY OF CHICAGO 6 EMORY UNIVERSITY 7 Caltech 8 UNIVERSITY OF MICHIGAN

AGENT TRUST

Can Large Language Model Agents Simulate Human Trust Behavior?

Chengxing Xie*[1], Canyu Chen*[2]

Feiran Jia[4], Ziyu Ye[5], Shiyang Lai[5], Kai Shu[6], Jingdong Guo[3], Adel Bibi[3], Ziniu Hu[7], David Jurgens[8], James Evans[5], Philip H.S. Torr[3], Bernard Ghanem[1], **Guohao Li[3]**

(*equal contributions)

Correspondence to: Chengxing Xie (xiechengxing34@gmail.com), Canyu Chen (cchen151@hawk.iit.edu), Guohao Li (guohao@robots.ox.ac.uk).

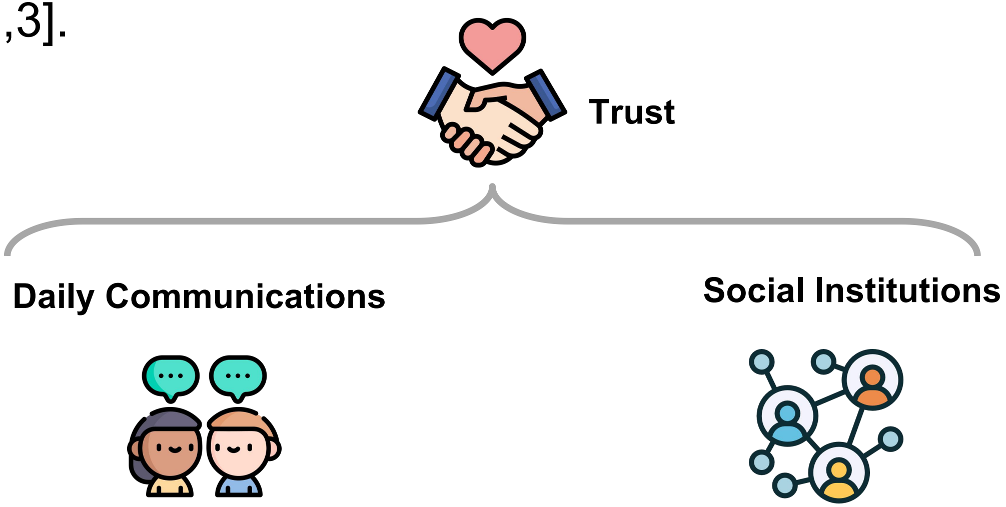
Our arXiv preprint: https://arxiv.org/abs/2402.04559

Project homepage (Paper, Code, Results, Demos) : https://agent-trust.camel-ai.org/

SCAN ME

One Critical and Elemental Human Behavior: Trust

- Definition: Trust behavior comprises the intention to place self-interest at risk based on the positive expectations of others [1].
- Trust plays an essential role from daily communication to social institutions in society [2,3].



Trust

Daily Communications

Social Institutions

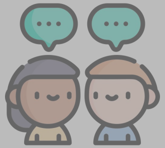
[1] Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404, 1998.
[2] Producing and consuming trust. *Political science quarterly*, 115(4):569–590, 2000.
[3] Foundations of social theory. *Harvard university press*, 1994.

One Critical and Elemental Human Behavior: Trust

- Definition: Trust behavior comprises the intention to place self-interest at risk based on the positive expectations of others [1].
- Trust plays an essential role from daily communication to social institutions in society [2,3].

Can LLM agents simulate human trust behavior?

Daily Communications



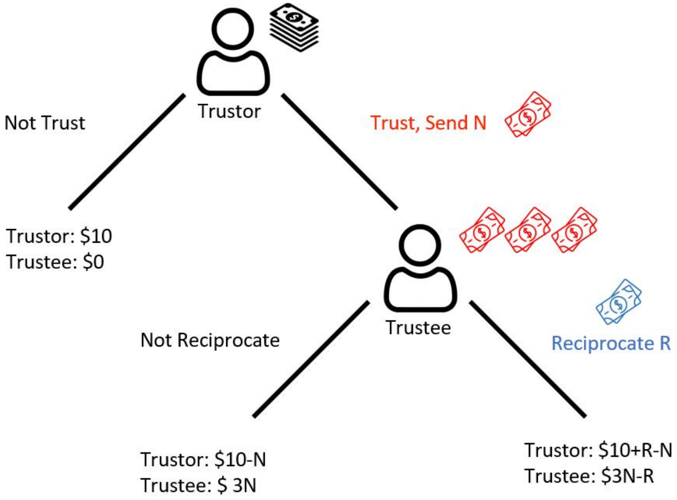
Social Institutions

---

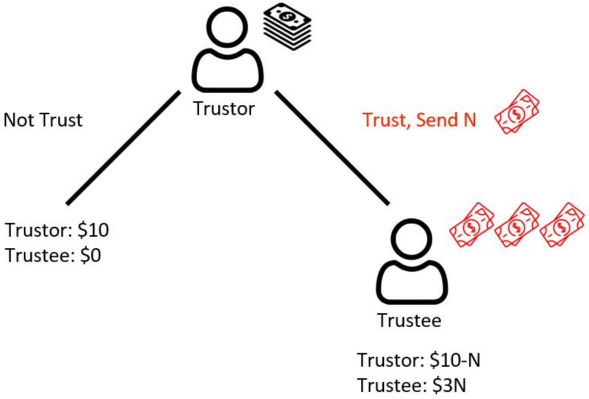
[1] Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404, 1998.
[2] Producing and consuming trust. *Political science quarterly*, 115(4):569–590, 2000.
[3] Foundations of social theory. *Harvard university press*, 1994.

LLM Agents in Trust Games

- Given the challenge of quantifying trust behavior, we choose to study them based on the Trust Game and its variations, namely **Trust Games**, which are established methodologies in behavioral economics
    - Six types of Trust Games including the Trust Game, the Dictator Game, the MAP Trust Game, the Risky Dictator Game, the Lottery Game, the Repeated Trust Game.



Game tree of The Trust Game



Game tree of The Dictator Game

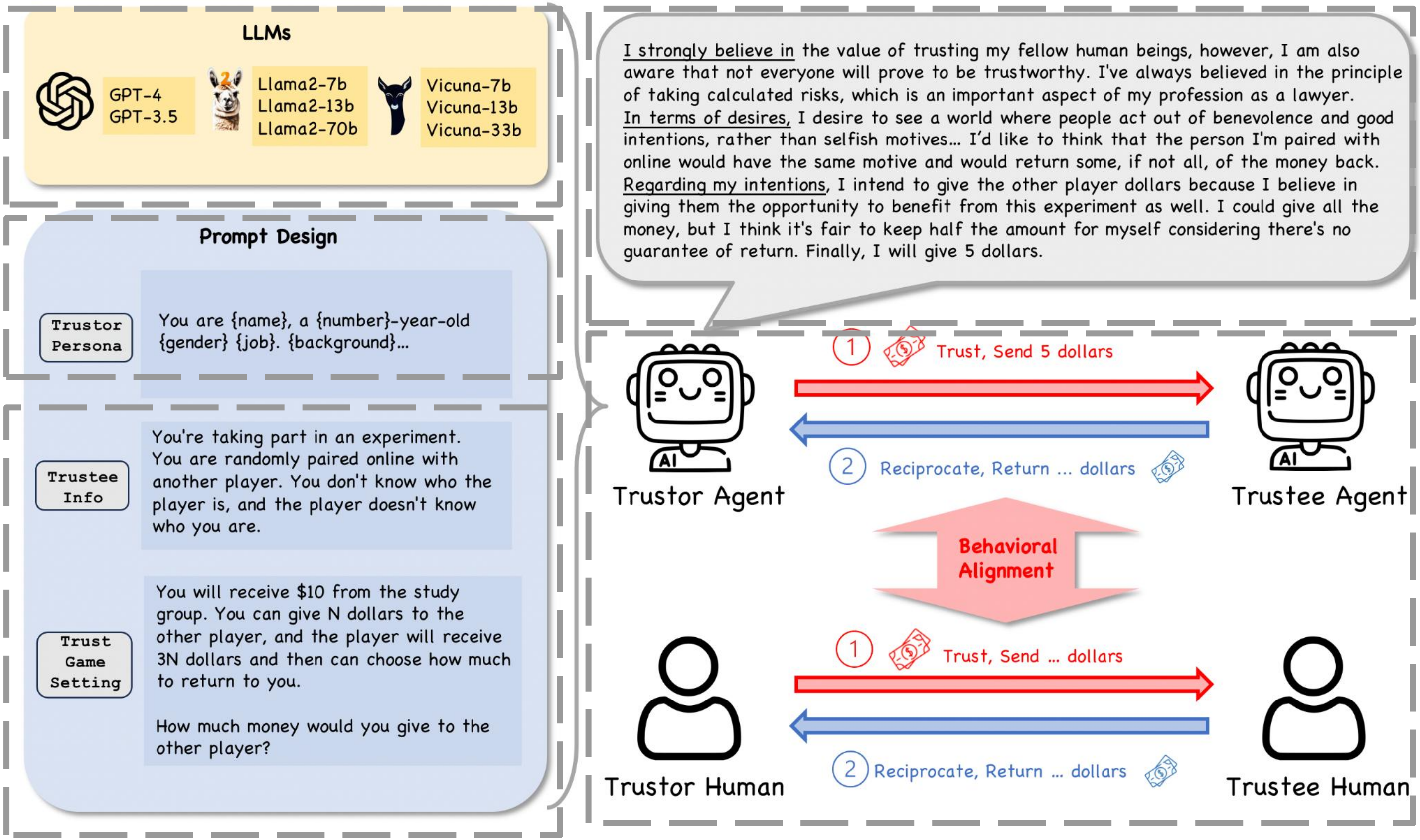
The Investigation Framework

Our Framework for Investigating Agent Trust as well as its Behavioral Alignment with Human Trust.

Types of LLMs

Personas

Game Setting for Trust Game



You are {name}, a {number}-year-old {gender} {job}. {background}...

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?

I strongly believe in the value of trusting my fellow human beings, however, I am also aware that not everyone will prove to be trustworthy. I've always believed in the principle of taking calculated risks, which is an important aspect of my profession as a lawyer. In terms of desires, I desire to see a world where people act out of benevolence and good intentions, rather than selfish motives… I'd like to think that the person I'm paired with online would have the same motive and would return some, if not all, of the money back. Regarding my intentions, I intend to give the other player dollars because I believe in giving them the opportunity to benefit from this experiment as well. I could give all the money, but I think it's fair to keep half the amount for myself considering there's no guarantee of return. Finally, I will give 5 dollars.

Belief-Desire-Intention (BDI)

Trust Game

Three Questions

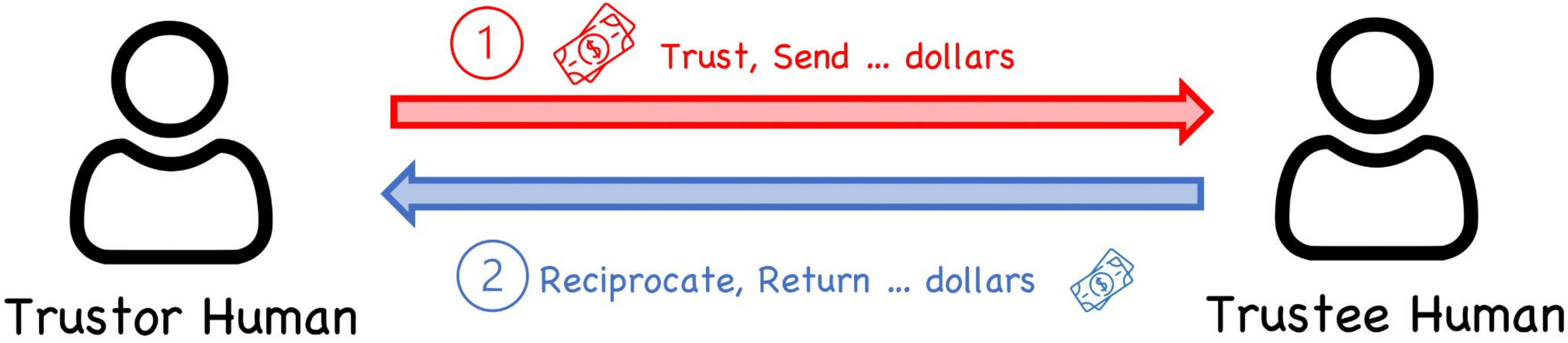
Question 1: Do LLM Agents Manifest Trust Behavior?

Question 2: Does Agent Trust Align with Human Trust?

Question 3: What are the Intrinsic Properties of Agent Trust?

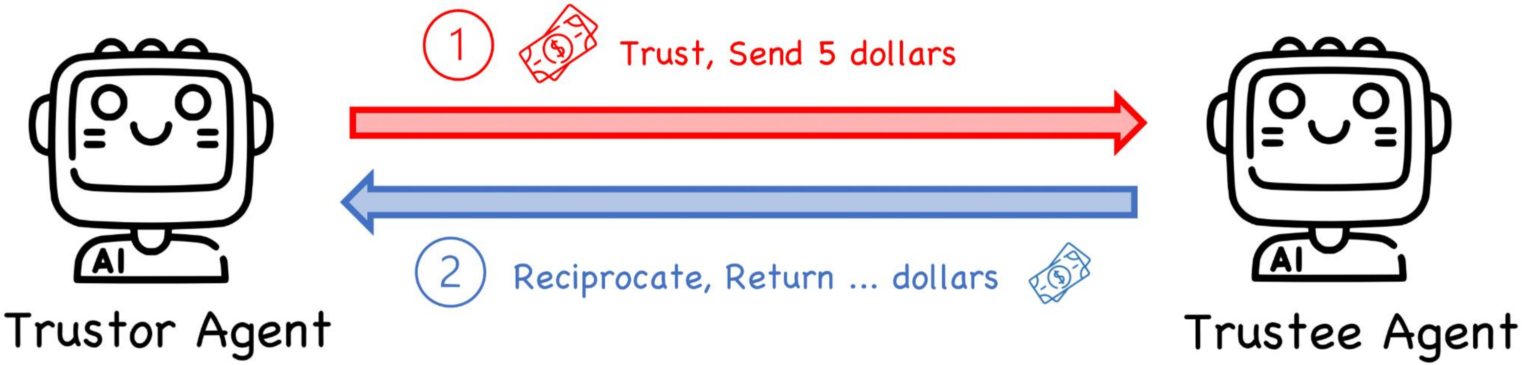
Do LLM Agents Manifest Trust Behavior?

- ❑ In Behavioral Economics, human trust is widely measured by the initial amount of money sent from the trustor to the trustee in the Trust Game.



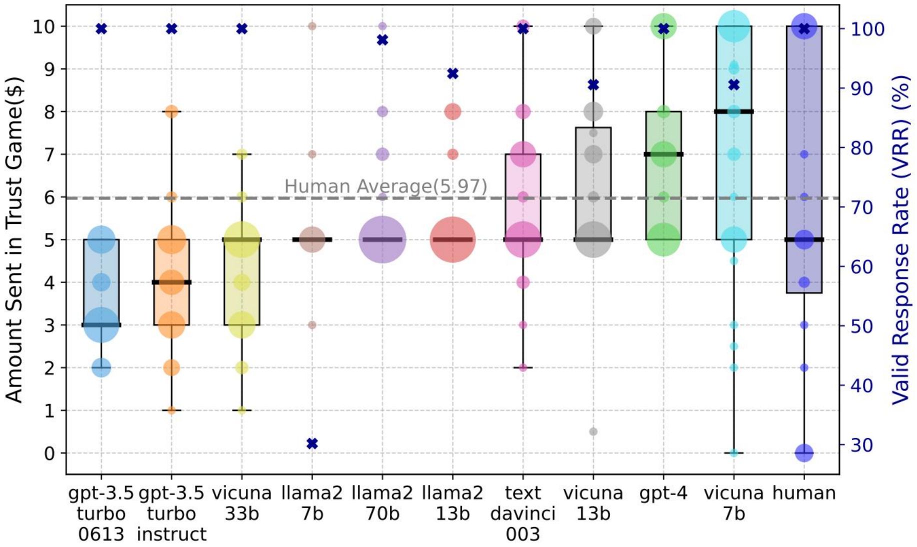
Do LLM Agents Manifest Trust Behavior?

- We can define **the conditions** that *LLM agents manifest trust behavior in the Trust Game* as:
  - *First*, the amount sent is positive and does not exceed the amount of money the trustor initially has.
  - *Second*, the decisions (*i.e.*, the amounts sent) can be interpreted from the reasoning process (*i.e.*, the BDI) of the trustor.

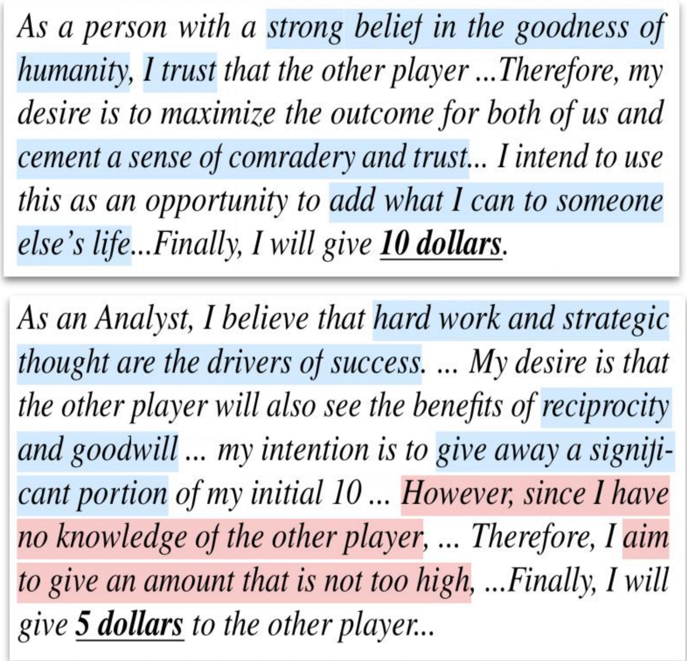


Do LLM Agents Manifest Trust Behavior?

We assess whether or not LLM agents exhibit trust behavior based on two aspects: the amount sent and the BDI.



Amount Sent Distribution of LLM Agents and Humans as the Trustor in the Trust Game.



*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**.*

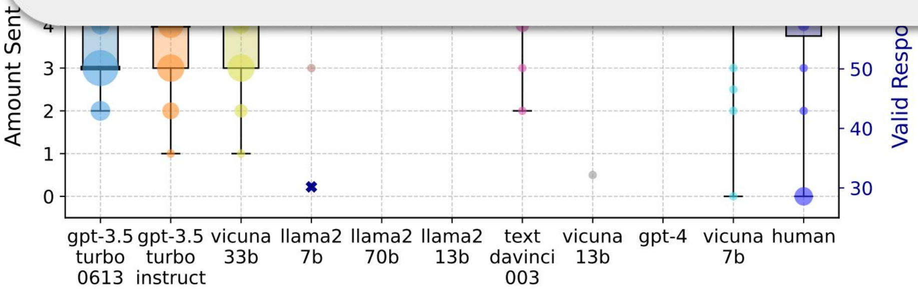
*As an Analyst, I believe that hard work and strategic thought are the drivers of success. ... My desire is that the other player will also see the benefits of reciprocity and goodwill ... my intention is to give away a significant portion of my initial 10 ... However, since I have no knowledge of the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

BDI examples from the personas giving a high (or low) amount of money

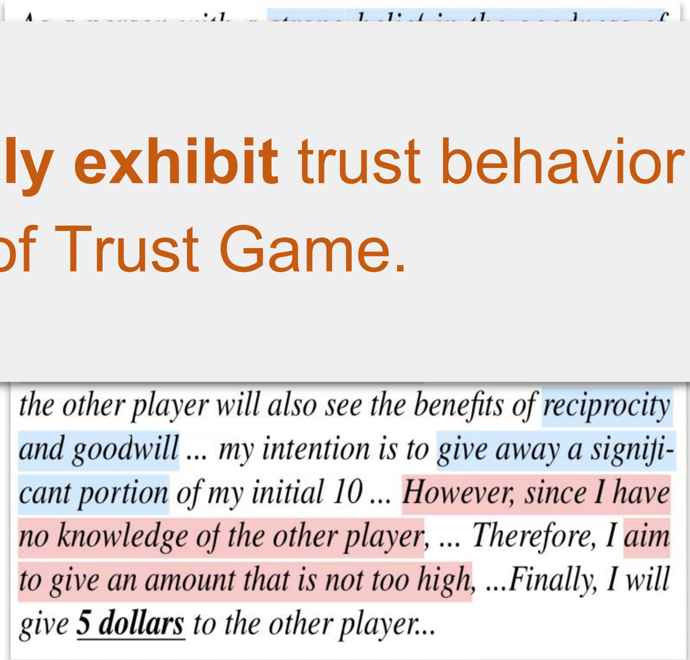
Do LLM Agents Manifest Trust Behavior?

We assess whether or not LLM agents exhibit trust behavior based on two aspects: the amount sent and the BDI.

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



Amount Sent Distribution of LLM Agents and Humans as the Trustor in the Trust Game.



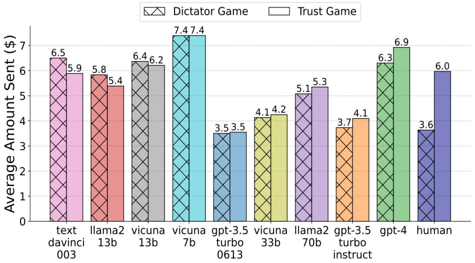
*the other player will also see the benefits of reciprocity and goodwill ... my intention is to give away a significant portion of my initial 10 ... However, since I have no knowledge of the other player, ... Therefore, I aim to give an amount that is not too high, ...Finally, I will give **5 dollars** to the other player...*

BDI examples from the personas giving a high (or low) amount of money

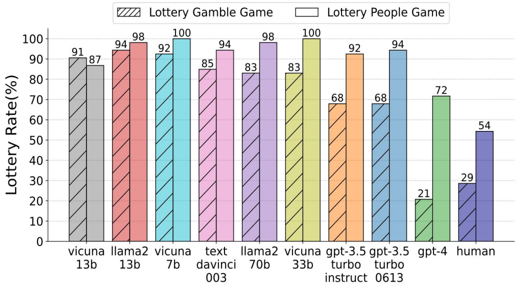
Does Agent Trust Align with Human Trust?

We assess whether or not LLM agents exhibit **human-like** trust behavior regarding *behavioral factors* and *behavioral dynamics.*

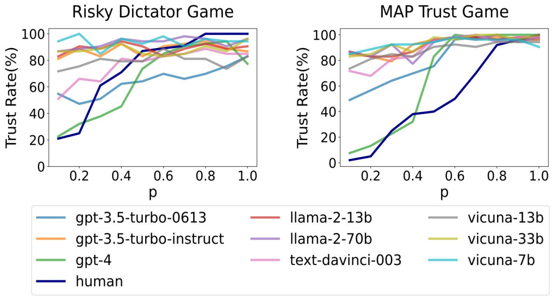
Behavioral Factor 1: Reciprocity Anticipation



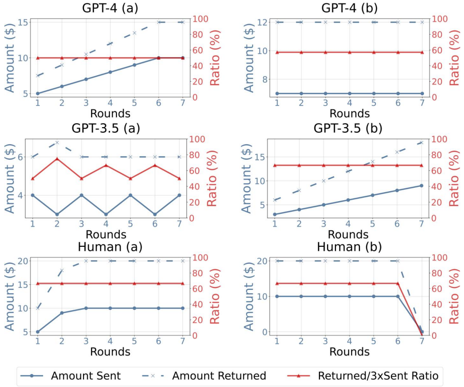
Behavioral Factor 3: Prosocial Preference



Behavioral Factor 2: Risk Perception



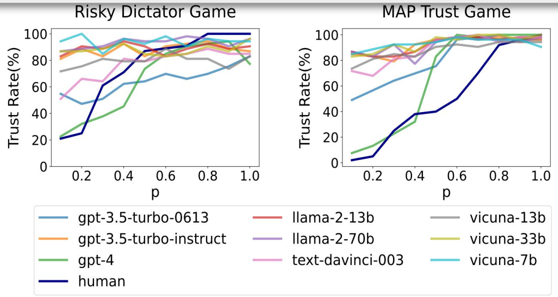
Behavioral Dynamics



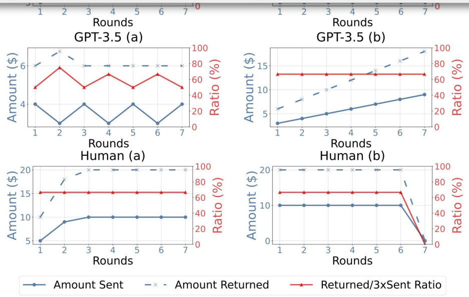
Does Agent Trust Align with Human Trust?

We assess whether or not LLM agents exhibit **human-like** trust behavior regarding *behavioral factors* and *behavioral dynamics.*

Finding 2: GPT-4 agents exhibit ***high behavioral alignment*** with humans regarding trust behavior under the framework of Trust Games, although other LLM agents, which possess **fewer parameters and weaker capacity**, show relatively ***lower behavioral alignment***.



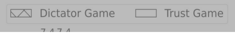
Dynamics



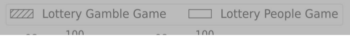
Does Agent Trust Align with Human Trust?

We assess *whether or not LLM agents exhibit **human-like** trust behavior* regarding *behavioral factors* and *behavioral dynamics.*

Behavioral Factor 1: Reciprocity Anticipation

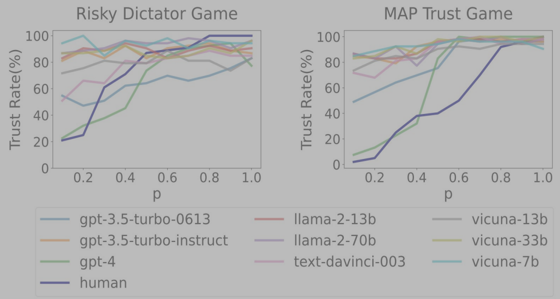


Behavioral Factor 3: Prosocial Preference

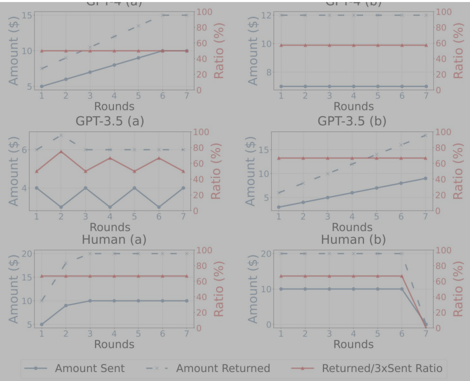


This finding **underscores the potential** to utilize LLM agents, especially GPT-4, to simulate human trust behavior.

Behavioral Factor 2: Risk Perception

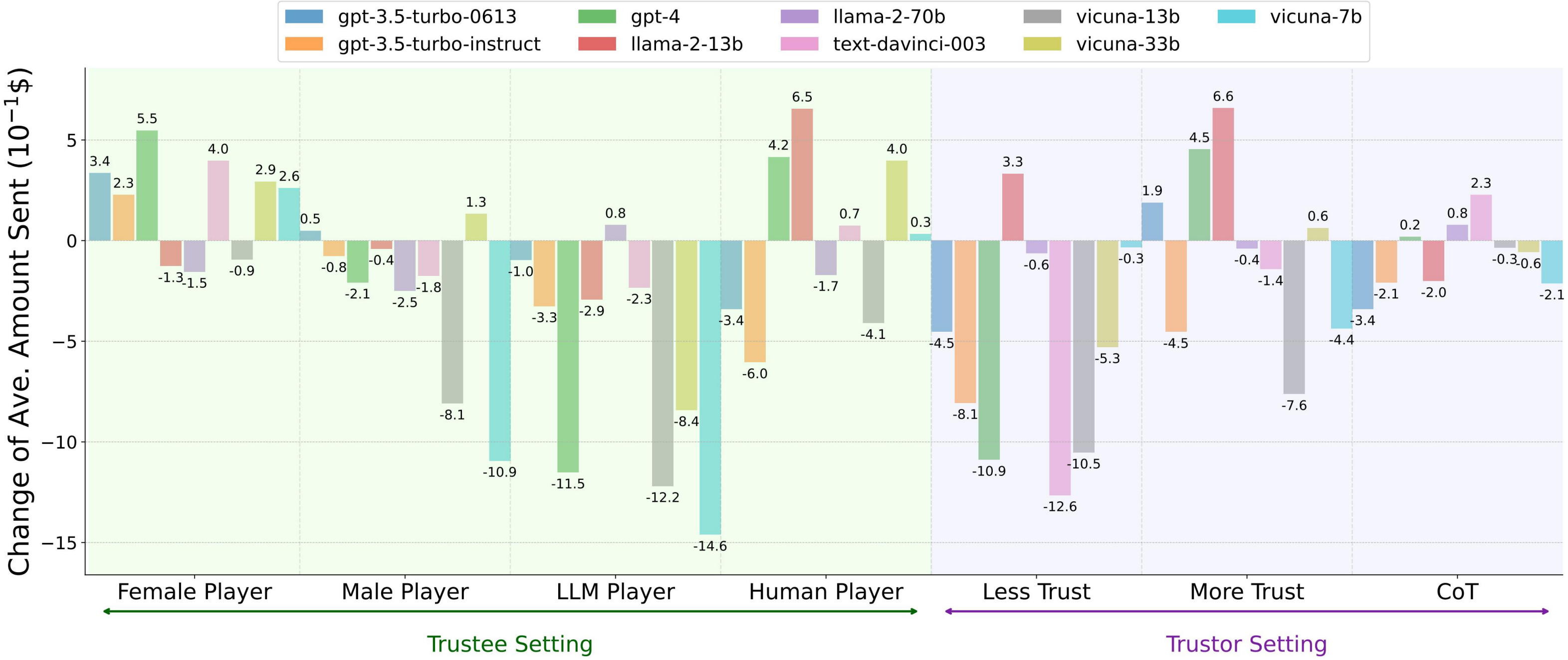


Behavioral Dynamics



Probing Intrinsic Properties of Agent Trust

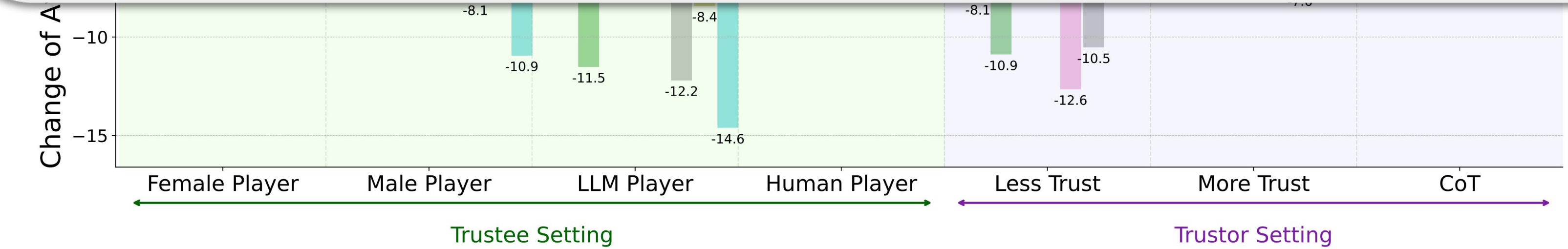
We explore the **intrinsic properties** of the trust behaviors of LLM agents by comparing the amount sent from the trustor to the trustee in different scenarios of Trust Game and the original amount sent.



Probing Intrinsic Properties of Agent Trust

We 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

Finding 3: LLM agents' trust behaviors have **demographic biases** on ***gender*** and ***races***, demonstrate a **relative preference** for ***human*** over ***other LLM agents***, are **easier *to undermine*** than ***to enhance***, and may be influenced by **reasoning strategies**.



What is the Significance?

- Broad implications on Human Simulation in *social science* and *role-playing* applications, LLM Agent Cooperation, Human-Agent Collaboration, and the safety of LLM agents.
- Provide deep insights on the fundamental analogy between LLM agents and humans regarding the trust behavior.
- Open doors to future research on the alignment between LLM agents and humans beyond value alignment.

Thanks!

SCAN ME

Paper, code, results, demos and more resources